SPIM

Habilitation à Diriger des Recherches

école doctorale sciences pour l'ingénieur et microtechniques

UNIVERSITÉ DE FRANCHE-COMTÉ

Energy Efficient Data Collection and Processing for Large-Scale Wireless Sensor Networks

Abdallah Makhoul

SPIM

Habilitation à Diriger des Recherches

école doctorale **sciences pour l'ingénieur et microtechniques**

UNIVERSITÉ DE FRANCHE-COMTÉ

HABILITATION À DIRIGER DES RECHERCHES

de l'Université de Franche-Comté

préparée au sein de l'Université de Franche-Comté

Spécialité : **Informatique**

présentée par

ABDALLAH MAKHOUL

Energy Efficient Data Collection and Processing for Large-Scale Wireless Sensor Networks

Soutenue publiquement le November 9, 2018 devant le Jury composé de :

PR. LYNDA MOKDAD	University of Paris-Est Val de Marne	Reviewer
PR. DJAMAL BENSLIMANE	University of Lyon 1	Reviewer
PR. YE-QIONG SONG	University of Lorraine	Reviewer
PR. HAMAMACHE KHEDDOUCI	University of Lyon 1	Examinator
PR. CONGDUC PHAM	University of Pau et Pays de l'Adour	Examinator
PR. JACQUES BAHI	University of Bourgogne Franche-Comté	Examinator
PR. RAPHAËL COUTURIER	University of Bourgogne Franche-Comté	Examinator

RÉSUMÉ

Collecte et Traitement de Données Massives dans les Réseaux de Capteurs à Large Échelle

Abdallah Makhoul
Université de Franche-Comté, 2018

Les recherches présentées dans ce document s'inscrivent dans le cadre des réseaux de capteurs sans fil et par conséquent de l'Internet des objets. Les réseaux de capteurs sans fil constituent l'une des technologies les plus cruciales pour le développement de l'internet du futur. Un réseau de capteurs peut être formé en déployant des capteurs spécifiques dans la zone d'intérêt afin de la surveiller. Ces réseaux s'insèrent dans de nombreux domaines d'application, comme la surveillance de l'environnement, le monde médical, l'industrie, l'aéronautique, etc. Beaucoup de ces applications exigent des réseaux de capteurs à large échelle, où un grand nombre de capteurs sont déployés dans une large zone géographique. Par conséquent, de nouveaux besoins sont créés pour comprendre et concevoir les systèmes, surtout que ces capteurs sont généralement très limités en termes de ressources. Ces limites s'expriment sous forme de contraintes de taille, de consommation d'énergie et de puissance de calcul. En effet, la gestion de données massives en temps réel, la prise de décision, la sécurité, etc., seront des composantes essentielles à prendre en compte pour assurer le bon fonctionnement et la meilleure qualité de surveillance. Dans le cadre des réseaux de capteurs à large-échelle, les recherches présentées dans ce document s'articulent autour de trois enjeux majeurs. Le premier porte sur la collecte et l'analyse de données massives, le deuxième sur l'agrégation et la fusion de données, et le troisième sur la sécurité et la survie de données.

Les réseaux de capteurs collectent des données de l'environnement et collaborent ensemble pour comprendre le phénomène surveillé. Comme chaque nœud est une source de données, et comme il peut être équipé d'un ou de plusieurs capteurs, de nombreuses données seront recueillies. Cependant, l'analyse des flux de données pour obtenir des informations et prendre des décisions appropriées, est l'un des défis de conception pour les réseaux de capteurs et les applications de surveillance. La gestion de données n'est pas une tâche facile, en particulier pour des capteurs ayant des ressources limitées. Un premier objectif de nos recherches consiste donc à réduire cette masse de données tout en conservant son intégrité. Par conséquent, l'énergie d'un nœud-capteur qui est consommée dans ces trois phases : la collecte, le traitement et la transmission de données, doit être bien gérée pour obtenir un équilibre entre la qualité des données et la consommation d'énergie. Plusieurs modèles pour la collecte de données ont été étudiés. Dans ce document, nous présentons des modèles d'acquisition de données permettant à chaque nœud d'adapter son taux d'échantillonnage à l'évolution dynamique de

l'environnement. Ces modèles permettent la réduction du sur-échantillonnage et par conséquent la réduction de la quantité d'énergie consommée. Une autre technique étudiée et présentée dans ce document est le mécanisme de double prédiction des séries temporelles. Un modèle de prédiction identique est partagé entre chaque nœud capteur et le Sink. Ce modèle est utilisé pour prédire les valeurs futures. Ainsi, au lieu de transmettre toutes les données collectées, un capteur ne transmet que les mesures qui s'écartent de la valeur prédite d'un seuil prédéfini par l'utilisateur.

Une technique efficace permettant la réduction de la taille de données est l'agrégation. En effet, les données produites par les capteurs voisins sont très corrélées spatialement et temporellement. Ceci peut engendrer la réception par l'utilisateur final des informations redondantes. Réduire la quantité de données redondantes transmises par les nœuds permet de réduire la consommation d'énergie et économiser de la mémoire dans le système et prépare les données pour la prise de décision. Dans la seconde partie de ce document, nous présentons trois techniques différentes pour l'agrégation de données dans les réseaux de capteurs. Le but principal est d'identifier tous les nœuds voisins qui génèrent des séries de données similaires. Deux couches d'agrégation sont proposées. La première au niveau des nœuds eux-mêmes et la deuxième au niveau des agrégateurs. Les trois méthodes proposées sont basées respectivement sur les fonctions de similarité, l'algorithme K-moyenne et les tests statistiques, et les fonctions de distance. Nous avons comparé ces différentes fonctions entre elles et nous avons proposé un modèle de filtrage par fréquence permettant d'optimiser le temps de calcul et d'améliorer la latence de données.

Dans la troisième partie nous nous intéressons à la fusion des informations provenant des différents capteurs pour l'aide à la décision. Il s'agit de traduire les règles et décisions définissant les événements lors de la surveillance (médicale, environnementale, etc.) et aussi de construire des liens de cause à effet, afin de proposer aux utilisateurs finaux une information complexe permettant une bonne compréhension du phénomène surveillé. Pour cela, en premier lieu nous avons étudié un modèle de fusion de données au niveau du coordinateur (ex. agrégateur). Ce modèle se base sur la logique floue, une matrice de décision et un système d'alerte précoce. Cette technique permet au coordinateur de prendre des décisions en fonction des données récoltées par les capteurs. Un autre modèle de fusion de données est présenté dans ce document dédié aux applications médicales. Ce modèle permet au coordinateur de calculer le risque de l'état du patient et de prendre en conséquence la décision convenable. Il est basé sur un système d'inférence floue ayant comme entrée le score agrégé de tous les signes vitaux et comme sortie le niveau du risque de l'état du patient. Dans cette approche, les décisions sont prises périodiquement et à chaque fois qu'un événement critique est détecté par le coordinateur.

La sécurisation des données au sein des réseaux de capteurs est une des préoccupations récurrentes dans ce domaine, et elle a connu de nombreux développements au cours de ces dernières années. Dans la dernière partie de ce document, nous étudions un modèle pour la survie de données. Il s'agit de garantir que l'information collectée restera accessible, et sera toujours présente dans le réseau lorsque la présence du puits est sporadique. Nous avons étudié un modèle pour la survie de l'information basé sur les modèles épidémiologiques de type SIR (Susceptible - Infected - Recovered), divisant une population d'individus en trois compartiments : ceux susceptibles d'être infectés, les individus malades, et ceux qui ont été soignés. Un des buts de ce genre de modèles est de s'assurer que le nombre d'infectés ne va pas ex-

ploser. Nous avons considéré que le réseau de capteurs est divisé en trois compartiments semblables de capteurs, et nous avons démontré que sous certaines conditions le nombre de capteurs informés (les «infectés») ne devient jamais nul.

Finalement, pour valider les approches proposées plusieurs expérimentations sur des réseaux de capteurs réels déployés au sein de notre laboratoire ont été réalisées. Une plate-forme de 30 nœuds Crossbow TesloB ont été déployés pour collecter des données environnementales de la zone surveillée (température, humidité, lumière).

MOTS-CLÉS : Réseaux de capteurs, algorithmes distribués, collecte et prédiction de données, analyse de la variance, agrégation et fusion de données, fonctions de similarité, partitionnement en k-moyennes, système d'inférence floue, survie de données, modèles épidémiologiques.

Contents

LIST OF FIGURES

List of Tables

LIST OF ALGORITHMS

DEDICATION

I dedicate this work

To the soul of my father, Geryes Makhoul. I miss him everyday and i know if he were here, he would be proud. I am sure, he is seeing me and offering me his ongoing support.

To my mother Eva, for her endless Love.

To my beloved wife Nelida who has been a constant source of support and encouragement. Thank you for your care, listening and understanding.

To my children Giorgio and Miléna, I wish to express my heartfelt and endless love to them.

To my sister Gretta and my brothers Elias and Peter, I love you all.

ACKNOWLEDGEMENTS

I would like to thank my colleagues professors Jacques Bahi, Raphaël Couturier and Christophe Guyeux for their support and help during all these years in Belfort to accomplish this HDR. They gave me many ideas and advices and the opportunity to co-supervise several phd students together.

I warmly thank professors Lynda Mokdad, Djamal Benslimane, Ye-Qiong Song, Hamamache Kheddouci and Congduc Pham, for accepting, despite their loaded schedules and busy agendas, to participate in the Habilitation jury and review my work.

During these last 9 years in Belfort, I had the chance to work with many PhD students, Maguy Medlej, Hassan Harb, Ibrahim Atoui, Carol Habib, Christian Salim, Anthony Tannoury, Nadine Bou Dargham, Gaby Bou Tayeh and Joseph Azar. We worked together on several issues related to Wireless Sensor Networks. I am really grateful to all of you.

I have been now working for over ten years in Belfort as an assistant professor in the AND team. I would like to thank all my AND team colleagues. It is hard to list all of them. I consider myself very fortunate for having a chance to work with you in such a good atmosphere. Thanks so much.

Finally, I am grateful to my lovely wife for her love and patience. Without her this work would have never been finished.

Thank you, everyone.

Introduction

This document summarizes my researches in the field of Wireless Sensor Networks (WSN). WSN is one of the most crucial technologies for the development of the Internet of things. A sensor network can be formed by deploying specific sensors in the area of interest to monitor it. These networks are used in many fields of application, such as environmental monitoring, health, industry, aeronautics, etc. Many of these applications require large-scale sensor networks, where a large number of sensors are deployed over a wide geographical area. In the context of large-scale sensor networks, the researches presented in this document focus on three major issues. The first deals with massive data collection and prediction, the second with data aggregation and fusion, and the third with security and data survival.

1. Research Context

The rapid proliferation of Wireless Sensor Networks (WSN) has given rise to various concepts that integrate the physical world with the virtual one. One of the most popular is the Internet of Things. A vision in which billions of smart objects are linked together, thus enabling anytime and any place connectivity for anything and not only for anyone. In the Internet of things, “things” are expected to become active and enabled to interact and communicate among themselves and with the environment by exchanging sensed data and information. Thus, they react autonomously to the real world events with or without direct human intervention. WSNs are considered as an integral and main part of the Internet of Things paradigm since they can provide a digital interface to the objects of the real world. In this future interconnected world, multiple sensors join the internet dynamically, and use it to exchange information all over the world in semantically interoperable ways [1].

A wireless sensor network is generally composed of a large number of nodes that cooperatively monitor and control the environment. Each node is equipped by sensors to detect physical phenomena like, temperature, humidity, vehicular movement, lightning condition, pressure, noise levels, etc. This enables large number of potential applications. Today, WSNs are used to monitor the physical conditions with their applications extended to health, automation, vehicular networks, industrial infrastructure, traffic, etc. [2]

A sensor node is of tiny size, consumes extremely low energy, and can be adaptive to the surrounded environment. It is composed of four hardware components: power unit, sensors, CPU and memory, and wireless transceiver. As these nodes are typically tiny devices, they can only be equipped with a limited power source. Furthermore, they have limited sensing and computation capacities and communication performance. Each sensor node has a zone of coverage for which it can reliably and accurately collect and transmit information. Sensor nodes can sense the environment and communicate the information through wireless links to a base station called ”Sink”.

WSNs may potentially have an impact on the interactions between humans and their environment by providing, processing and delivering any sort of information or command. However, due to limited storage capacity and power of sensor nodes, many challenges must be addressed in order to setup a workable sensor network. Energy conservation is of prime-concern and the most challenging problem in designing sensor networks. Besides, other related main challenges, include massive data collection, data reduction, synchronization, QoS, security, topology and architecture, etc. [3].

2. Challenges

Wireless sensor networks face some issues mainly related to energy conservation, energy efficient data management, clustering and security.

Energy Consumption

In order to monitor the environment and ensure high QoS, while maintaining the network connectivity, energy of nodes need to be conserved to extend the network lifetime. A sensor node consumes energy in three states, data collection, data processing, and data communication. With the scale of sensor networks increasing, changing or recharging batteries is no longer applicable. Thus energy saving becomes a key issue in large-scale WSN, especially, in some harsh environment, such as rainforest, volcano, etc. Therefore, the most crucial research challenge for the WSN researchers is to design, develop and implement energy efficient hardware and software solutions and energy supply technology for WSNs.

Massive data management

Usually hundreds or thousands of sensor nodes are randomly deployed to collect environmental data for a region of interest. This makes WSNs one of the big data producers. This fact has been supported by the report of ORACLE [4] where some examples of applications generating big sensor data were provided. In addition, the authors in [5, 6] give many real WSN applications where the scale of the sensory data has already exceeds several petabytes (PB) annually. However, such big data applications raise two problems: high energy consumption and complex data analysis. First, the sensing of big data volume leads to a great waste of sensors energy, which is usually limited and not rechargeable, thus decreases the network lifetime. Second, it is a complicated mission for data scientists when dealing with a big amount of sensed data, that mostly contain a high redundancy level, to make the right decisions. Therefore, the way these data are manipulated (collected, processed, delivered, and assessed) by the sensor nodes is a fundamental issue.

Data Acquisition and collection

Data Acquisition is the first phase in the sensory data life cycle. In this phase, the sensors sample the data/measures from the monitoring physical world. In order to keep the networks operating for long time, adaptive sampling approach to periodic data collection constitutes a fundamental mechanism for energy optimization. The key idea behind this approach is to allow each sensor node to adapt its sampling rates to the physical changing dynamics. In this way, over-sampling can be minimized and power efficiency of the overall network system can be further improved. Another efficient data collection technique for reducing the energy consumption in WSN is online data prediction models. They enable both the node and the Sink to predict sensor readings simultaneously. Thus, it is only required that the node sends the measurements that deviate from the prediction by a user predefined error threshold. The main challenges of such approximate data collection techniques are how to maintain the local and global models valid throughout the network lifetime. On the other side, the main advantage of these models is that they only transmit a partial amount of data to the sink, since a number of sensory values are either not sensed or can be predicted.

Data Aggregation

In large scale WSN, sensory data reported by the neighbouring nodes has some degree of redundancy. For instance, in environmental monitoring applications, it is generally the case that neighboring nodes monitoring a specific feature typically collect identical or similar values. Therefore, transmitting raw data separately in each sensor node consumes more energy and bandwidth and shortens the network lifetime. To reduce communication costs and energy consumption, data aggregation techniques have been introduced. They consist on removing or reducing nodes redundant sensor data and avoid forwarding multiple copies of information. It is noted that the energy consumed in transmission is much greater than that in processing in WSN. Therefore, an important issue in large scale WSN is to remove large quantities of redundant information, so as to minimize the amount of transmission and maximize the network lifetime.

Efficient Data Fusion

Wireless sensor networks are usually deployed to collect and process data from the environment in order to have a better understanding of the monitored condition. A fundamental issue in WSN is the way to process the gathered data. Early detection, alarming and initiation of emergency measures are key steps to avoid major environment disasters. In this case, data fusion arises as an efficient method to increase the significance of the collected data and the alarming efficiency. Data fusion can be defined as the combination of data gathered by multiple sensor nodes. By exploiting the synergy among these data, information fusion techniques can lead to an improved information with greater relevance.

Security

WSNs are often deployed in public or otherwise untrusted and even hostile environments, which prompts a number of security issues (e.g., key management, privacy, access control, authentication, *etc.*). Then, if security is necessary in other (e.g., wired or MANET) types of networks, it is much more so in sensor networks. Actually, it is one of the most popular research topic in this field and many advances have been reported on in recent years. In WSNs, it is essential for each sensor node and the sink to be able to verify that the data received was really sent by a trusted node and not by an adversary that sent false data. From another side, data integrity should be preserved and accurate data must reach at user end.

We can also notice the importance of a cooperative secure data aggregation in sensor networks. In other terms, after the data gathering and during transmissions to the base station, each node along the routing path cooperatively integrates and secures the fragments messages. Therefore, secure data aggregation protocols require sensor nodes to encrypt or authenticate any sensed data prior to its transmission, implement data aggregation at every intermediate node (without decryption), and prefer data to be decrypted by the sink so that energy efficiency is maximized.

Another important issue is data survivability in unattended WSN (UWSN) which are characterized by the sporadic presence of the sink. In such networks, nodes collect data from the area of interest, and then they try to upload all the stored data when the sink comes around. Due to the absence of a direct and alive connection with the sink, these networks are more subject to malicious attacks than traditional WSNs. Therefore, the critical issue for UWSNs is how to maximize information survivability which consists on preserving data for a long period of time in the face of attacks.

3. Main Contributions

In this section, we present a summary of our contributions to the challenges introduced above, then, later in next chapters we will detail some of them.

Massive data Management

Mass data are usually collected and processed in large-scale WSNs, and this will affect the lifetime of sensor nodes and the performance of network. Therefore, managing this huge amount of collected data is not an easy task, especially for sensors with limited energy and computational resources. One of our main objectives is to reduce this mass of data while maintaining its integrity. The energy of a sensor node which is consumed in three phases: data collection, processing and transmission, must be well managed to achieve a balance between data quality and energy consumption. In the literature, one can find various data reduction approaches proposed for WSNs. Some works are based on in-network processing and using algorithms like least mean square and Kalman filter. Other works are based on stochastic approaches, time-series forecasting, heuristics and algorithms. Moreover, data compression methods have been applied in sensor networks in order to reduce the size of data transmitted in the network by involving encoding at nodes and decoding at the sink. Although these approaches allow efficient data reduction, however they present several disadvantages. They are almost complex, sometimes they generate communication overhead, and the sink may need further transmissions to detect failures. In our researches, we designed and studied several protocols and models for data management in WSN in order to reduce the large volumes of collected data, optimize energy consumption and facilitate knowledge extraction.

Data Acquisition and prediction (Chapter 1)

Adaptive sampling techniques are very promising, because of their efficiency to optimize energy consumption and the network overload. However, most of the previous proposed solutions are implemented in a centralized manner that requires rather huge computations and communications. Other existing methods are limited to only space correlation and based on grouping nodes into clusters. In our research, we proposed distributed adaptive sampling algorithms which are based on the sensed data variation. We study the dependence of measurements variance while taking into account the residual energy that varies over time. We exploit statistical tests based on one-way ANOVA model. Moreover, we take into account the application criticality and propose a model that dynamically defines multiple levels of sampling rate corresponding to how many samples are captured per unit of time. The final goal is to provide the necessary algorithmic support for environmental surveillance applications to express their objectives [7, 8, 9]. Second contribution is about data prediction. We exploit the fact that sensor data changes smoothly over time, therefore we use linear interpolation to predict future readings. Moreover, an algorithm is applied to dynamically adapt the interpolation line with the real readings curve. We also coupled this technique with a data reconstruction algorithm, that exploits both temporal smoothness and spatial correlation among different sensed features in order to estimate missing data [10]. We have evaluated these techniques on real-world data sets measuring different environmental features such as temperature, humidity, light, and voltage collected at our laboratory. For instance, we have successfully reduced data transmission up to 99.7% for temperature data, while maintaining an accuracy of 0.1 degree Celsius.

Data Aggregation (Chapter 2)

To study data aggregation in large scale sensor networks we considered that the nodes are organized into clusters and we proposed two layers of aggregation. A first in-sensor process layer is done by the nodes themselves. Instead of sending raw data, each sensor node reduces redundancies from the collected data before transmitting it to the cluster head (CH) for a second layer of aggregation. At the level of CH, our objective is to identify neighbouring nodes generating similar data sets. We studied three different techniques and we compared them together and with existing data aggregation methods. In the first one [11, 12], we investigate the problem of finding all pairs of nodes generating similar data sets such that similarity between each pair of sets is above a threshold t. We proposed a new frequency filtering approach and several optimizations using sets similarity functions to solve this problem. In the second contribution [13], we studied a new clustering method to handle the spatial similarity between node readings. Once the CH receives all data sets, it applies an enhanced K-means algorithm based on one-way ANOVA model to identify nodes generating identical data and to aggregate these sets before sending them to the sink. The third contribution [14] exploits the distance based functions (e.g. Euclidean, Camberra, Cosine, etc.) to find near data sets with the aim to eliminate redundancies and reduce the huge amount of data transmitted over the network. The obtained results show the efficiency of our methods in reducing redundant data, energy consumption and data latency.

Multisensor Data Fusion (Chapter 3)

Another challenge in WSN is data fusion which enables combining information from several sensor nodes to represent the global situation of the monitored process leading consequently to take a right decision. Our first contribution [15, 16] aims to obtain information of greater quality and make accurate decisions about the situation of the monitored condition based on the collected data. Our data fusion scheme uses Fuzzy set theory. The raw data received during consecutive periods are aggregated using fuzzification procedures. Then, the decision having the closest feature values to the aggregated data set is selected from a decision matrix put by experts. Our data fusion approaches have been tested in the e-health and wireless body sensor networks context. Second data fusion contribution [17] consists on a multisensor data fusion approach enabling the determination of the patient's risk level for health assessment. This assessment will be performed based on the collected measurements of the vital signs. The proposed approach uses fuzzy sets to deal with uncertainties and a fuzzy inference system to map the aggregate score of vital signs to the patient's risk level. The proposed approaches are compared with other existing works and validated by a healthcare expert.

Security - Data survivability (Chapter 4)

One of our main contributions dedicated to the security in WSN is data survivability in unattended WSN (UWSN) where the presence of the sink is sporadic [18] and in the internet of things [19]. In this scenario, sensor nodes collect and store data locally and try to upload all the information once the sink becomes available. We focus on non-cryptographic approaches for data survivability. We proposed an epidemic-domain inspired approach to model the information survivability in UWSN. The model we studied is based on both SIR (Susceptible - Infected - Recovered) and SIS (Susceptible- Infected - Susceptible) models. A node is susceptible to a data item when it is online and functioning normally; it can receive the information that must survive. Our novelty comparing to existing works is that we study arbitrary dynamic network topologies instead of static networks. Furthermore, we provided a fully distributed algorithm which supports/covers different epidemic models. The aim of this algorithm is to ensure data survivability in UWSN by maintaining a subset of safe nodes in working state (not idle) while replacing/locking the attacked ones when needed.

4. Document Organization

The remainder of this document is divided into five chapters. After having introduced our main contributions, Chapter 1 details our obtained results related to data acquisition and prediction. Then, Chapter 2 focuses on our data aggregation techniques for large scale sensor networks. Chapter 3 describes our multi-sensor data fusion contributions. In Chapter 4 we present our approach for data survivability in unattended sensor networks. Finally, Chapter 5 summarizes our conclusions and details our main perspectives.

1

Data Collection and Prediction in WSN

Data reduction is an effective technique for energy saving in wireless sensor networks. It consists in reducing sensing and transmitting data while conserving high quality of service. In this chapter, we study data collection and prediction in order to increase the network lifetime and to reduce the huge amount of the collected data. First, we propose a Dual Prediction Mechanism (DPM) while taking into account the data loss. In a second step, we propose an adaptive sampling algorithm allowing each sensor node to adapt its sampling rate to the physical changing dynamics. Finally, we combine these two approaches allowing us to significantly decrease energy consumption and extend the whole network lifetime. Our study was evaluated on real-world data sets collected at our laboratory and compared to recent data reduction approaches. The results were promising in quality of the replicated measurements and transmission reduction.

1.1/ Introduction

A Wireless Sensor Network (WSN) is composed of a large number of small and low-cost devices called sensor nodes. These nodes collect and transfer data to a central workstation also known as Sink and they are applicable in a wide range of monitoring applications. However, sensor devices have a limitation in memory, energy, and processing capabilities. Therefore, several approaches have been proposed to reduce the energy consumption of these nodes. Since radio communication is the dominant factor of energy consumption in WSN, the most effective approach is the reduction of data transmission between the nodes and the sink.

One of the most commonly used technique to reduce radio communication is the dual prediction mechanism [20, 21, 22, 23]. An identical prediction model is shared between each node and the Sink. This model is used to forecast future values. Thus, instead of transmitting all the collected data, a sensor transmits only the measurements that deviate from the predicted value by a threshold predefined by the user. Therefore, if the Sink does not receive any measurement at a given time, it acknowledges that the model's prediction is within the error budget.

Another efficient data reduction technique used in WSN is adapting sampling. Indeed, due to the nature of WSNs, sensor data tend to change smoothly over time and it contains a significant chunk of redundant information. Therefore, to reduce the number of sampled data some researchers proposed several adaptive sampling techniques [24, 25, 26, 9, 8, 7] that dynamically increase or decrease the sampling rate of a sensor according to the level of variance between collected data over a certain period of time. In this way, the oversampling can be minimized and the power efficiency of the overall network system can be further improved. Thus, the sensing activity is reduced which in turn leads to a reduction in the transmission activity and the energy consumption.

Merging both adaptive sampling and dual prediction based transmission reduction into a single mechanism, can reduce energy consumption significantly compared with the approaches relying on either one of them.

In this chapter we present a complete framework for data reduction in sensor networks. Thus, allowing the sensor node to optimally use its allocated energy resources and extending the overall lifetime of the network, while preserving the quality of the collected data. We propose:

- a data transmission reduction algorithm that reduces the amount of data reported to the sink using a dual prediction model. In contrast to other similar techniques our model is light in term of computational cost and requires a very small memory footprint, yet it is robust and efficient.
- an efficient adaptive model of data collection dedicated to WSN. The main idea behind this approach is to allow each sensor node to adapt its sampling rate to the physical changing dynamics.
- a mechanism for coupling our transmission reduction algorithm with the adaptive sampling technique. Enabling the sensor to collect fewer measurements which in turn increases the efficiency of the transmission reduction algorithm and reduces the amount of energy consumed by the sensing activity.

The remainder of this chapter is organized as follows. In Section 1.2 the state of the art related to data transmission reduction, adaptive sampling and approximate replication of sensor data is briefly presented. In Section 1.3 our novel dual prediction based transmission reduction method is introduced. In Section 1.4 the adaptive sampling method is explained. Section 1.5 explains how the adaptive sampling and transmission reduction techniques can be merged together. The obtained experimental results are shown in Section 1.6. Finally, this chapter is concluded in Section 1.7.

1.2/ State of the art

In the literature, one can find various data reduction approaches based on in-network processing (adaptive sapling, data aggregation, etc.), data compression or data prediction methods.

1.2.1/ Data Compression

Data compression can be applied in WSN in order to reduce the size of data transmitted in the network by involving encoding at nodes and decoding at the sink [27, 28, 29]. For instance, [30] exploits raw signal processing and signal reconstruction to develop a reordering algorithm that resorts the sensor nodes at the Sink. This method enhances the sparsity of the signal by reducing the number of measurements needed for its reconstruction, consequently resulting in a low compression sampling rate that in turn scales down irrelevant communication traffic. The authors in [31] proposed a cluster-based quality-aware adaptive data compression scheme, which takes into consideration the application's data quality and it also limits information loss by using adaptive clustering and novel coding algorithm. Although these approaches allow efficient data reduction, however they present several disadvantages. They are almost complex, sometimes they generate communication overhead, and the sink may need some transmissions to detect failures.

1.2.2/ Adaptive Sampling

The main goal of an adaptive sampling approach is to make the rate of sensing dynamic and adaptable; if the sensor node can adapt its sampling rates to the changing dynamics of the condition or process, over-sampling can be minimized and the computational load at the sink will be more flexible.

The authors in [32] propose an energy-efficient adaptive sampling mechanism which uses spatio-temporal correlation among sensor nodes and their readings. The main idea is to carefully select a dynamically changing subset of sensor nodes to sample and transmit their data. In [33], a machine learning architecture for context awareness is proposed. It is designed to balance the sampling rates (and hence energy consumption) of individual sensors with the significance of the input from that sensor. In [34], the authors propose an Adaptive Sampling Approach to Data Collection (ASAP) which splits the network into clusters. A cluster formation phase is performed to elect cluster heads and to select which nodes belong to a given cluster. The metrics used to group nodes within the same cluster include the similarity of sensor readings and the hop count. Then, not all nodes in a cluster are required to sample the environment. [35] proposes a TA-PDC-MAC protocol, a traffic adaptive periodic data collection MAC which is designed in a TDMA fashion. This work is designed in the way that it assigns the time slots for nodes activity due to their sampling rates in a collision avoidance manner. In [26] the sampling rate of the sensor node is adapted by taking into consideration both the system and the application context levels. For instance, the availability of the energy for harvesting represents the system context. This availability is the criteria used to set the maximum sampling rate for the node. The application context is represented by the user request, where feedback from the system executing specific rules of user or field scientist is used to set the rates of sensor node sampling in optimal way.

Adaptive sampling techniques are very promising, because of their efficiency to optimize energy consumption and the network overload. However, most of the previous proposed solutions are implemented in a centralized manner that requires huge computations and communications. Other existing methods are limited to only space correlation and are based on grouping nodes into clusters. In this chapter we propose an adaptive sampling algorithm based on the dependence of conditional variance on measurements, e.g. one-way ANOVA model and statistical tests. In parallel, we provide a multiple levels adaptive model that takes into account the application criticality. It defines dynamically multiple levels of sampling rate corresponding to how many samples are captured per unit of time (or period).

1.2.3/ Prediction-based Data Reduction

Several approaches for a prediction-based data transmission reduction have been proposed in the literature. In the following, we list and discuss the most common techniques.

In [23] the authors proposed a Derivative Based Prediction model (DBP) that is computed based on a learning window, containing m data points. The model is linear, computed as the slope d of the segment connecting the average values over the first and last l edge points at the beginning and end of the learning window. The authors in [36] proposed a prediction model that is based on the Kalman filter. The same instance of this model is built by the Sink using historical data reported by the sensor and then is shared with the latter. The same model on both ends simultaneously performs linear predictions for future readings, enabling the sensor to transmit a measurement to the Sink only when the prediction is not accurate. The dual Kalman filter method requires a priori knowledge and statistical data on the environment being monitored, in order to build the model. Therefore, the authors in [37] proposed a dual prediction mechanism that is based on Least Mean Square (LMS) adaptive filter. LMS lends itself to be compact, light and requires no priory statistical knowledge of the data. Thus, it makes the prediction model more stable and adaptable with changes. The authors in [38] proposed to combine the LMS and Recursive Least Squares (RLS) adaptive filters in a single prediction model. Since the latter is able to achieve faster convergence and produces a prediction model that is more stable, RLS is used to build the prediction model. Once this model is built, the parameters are then passed to a LMS adaptive filter to perform predictions. The reason for this switch is that LMS has lower complexity, thus it suits the energy constraints of the sensor better than RLS. In [39] the authors proposed a dual prediction model that is based on the Hierarchical Least Mean Squares (HLMS) adaptive filter. The HLMS is a multi-level LMS filter, that makes a trade-off between increasing the complexity of LMS filter and having a better prediction filter.

The speed and the success of convergence of the adaptive filters is conditioned by some predefined parameters such as the ”step size”. A small alteration in these parameters can heavily affect the performance of these filters, and choosing an optimal value is not feasible most of the time, since it requires a training phase.

None of the aforementioned data reduction approaches provides a fully efficient scheme that is able to maintain a reliable and loss free communication link between the node and the Sink. Therefore, we study in this chapter a Fault Tolerant Data Transmission Reduction technique that estimates data without any prior knowledge about the statistical properties of the sensed data, nor a set of global parameters that control the performance of the prediction model. It also lends itself to be light, robust, and requires a very small memory space. Moreover, we have adopted and adapted the data reconstruction method proposed in [40] and we integrated it into our transmission reduction technique through a mechanism that can identify and flag missing data. We developed our technique with several goals in mind: better prediction accuracy, less energy consumption, less computational cost, and minimum prediction delay.

1.2.4/ Adaptive Sampling and Data prediction

Although there has been a large number of recent works on sensor networks, only a one recent work explicitly deals with combining adaptive sampling and data prediction in WSN [41]. The authors in [41] proposed a technique named Dual Prediction with Cubic adaptive sampling (DPCAS) that combines an exponential time series predictive model with a TCP CUBIC congestion adaptive sampling technique. It Enables the sensor node to reduce its sampling rate based on the produced prediction error. Moreover, measurements are transmitted to the sink only when a significant change in readings occur. The whole data set is then reproduced on the sink by interpolating the received measurements.

In this chapter we study also the combination of our adaptive sampling technique with our dual prediction based forecasting model that is free from any parameters limiting its performance and which only requires two measurements to be built and one measurement to be updated. Targeting at the same time the two most energy consuming activities in WSNs, in order to preserve as much energy as possible. We compare our approach with the technique proposed in [41].

In the following sections, data transmission reduction and adaptive sampling techniques are explained. Then an algorithm merging these two approaches together is presented.

1.3/ Dual Prediction Model (DPM) for transmission reduction

Let us first describe our data transmission reduction method, which aims to minimize the amount of data reported by each sensor to the sink.

At first, the sensor sends to the sink the two first readings at time t_0 and t_1. The sink stores in its memory the last value received which we will be referring to as "L" and the time when it is received t_L . Then, they both compute simultaneously the difference between these two measurements which will be referred to as d as shown in Eq. 1.1.

$$d[0] = x_1 - x_0 \tag{1.1}$$

When $d[0]$ is calculated, the sensor stops reporting readings to the sink. They both switch to the prediction phase and start predicting the value of the next reading $\widehat{x}_k$ at time t_k, by adding "$d[0]$" to the predicted value at a previous time tick $\widehat{x}_{k-1}$. d represents the change rate in readings since the last adaptation. Therefore, by adding d to the previous prediction to forecast the current one, we consider that the readings will keep changing following the same rate, which is based on the fact that time series data change smoothly over time. Moreover, such data are affected by three main factors including a cyclic factor, where they tend to follow the same cycle of increase, slow down and then decrease during the monitoring period. Therefore, d is multiplied by a rectification value α which can range between 0 and 1 in order to harmonize the prediction line with the real data curve. Hence, the predicted value at a time t_k is calculated according to the new Equation 1.2 provided below.

$$\widehat{x}_k = \widehat{x}_{k-1} + d[k] * \alpha \tag{1.2}$$

Once the estimation is calculated, the sensor compares the predicted value $\widehat{x}_k$ to the real sensed reading x_k. If it respects the error budget $emax$ defined by the user, the real reading is discarded. If the sink does not receive any value at a given time, it considers that the estimated value is within the error budget. Otherwise, if the estimation exceeds the error threshold, the sensor discards it and sends to the sink the real reading. When the Sink receives it, they both recalculate the value of d, by subtracting L from the current reading and dividing the result by the time tick difference between the previously received reading and the current one, as shown in Equation 1.3 below, and the value of L is replaced by the last received reading x_k.

$$d[k] = \frac{x_k - L}{t_k - t_L}. \tag{1.3}$$

1.3.1/ Updating α

As mentioned previously, the rectification value α is used to adjust the change rate d in order to adapt itself better with the upcoming data and extend the model's prediction horizon. In this section, we will explain how α is calculated automatically and in real time.

During the model update phase, both the sink and the node calculates a value called Accuracy Factor (AF) that reflects the accuracy of the prediction model. It is calculated by subtracting the value of the received measurement x_l at time t_l from its corresponding predicted value $\widehat{x_l}$ and dividing the results by the number of successful predictions that preceded the transmission (also known as prediction horizon). The smaller is the absolute value of AF the more accurate is the prediction model. Afterward, both the sensor and the sink calculate the percentage (P) of how much the value of AF is compared with $emax$.

In order to adjust the inclination angle of the linear prediction line, we consider that α should be $P\%$ less or more than it currently is. If the error $(\widehat{x_l} - x_l)$ is negative, this means the linear prediction line (or d) is decreasing faster than it should be. Thus, this decrease should be slowed down by increasing α by $P\%$. Otherwise, if the error is positive, this means that d is increasing faster than it should, thus it must be slowed down by reducing α by $P\%$. By doing so we aim to minimize AF for the next prediction phase, by assuming that it will remain similar or very close to the current one.

Three possible situations may occur when updating α:

- AF is very small compared to emax (e.g. 10% of emax): this means that the error is very small, and α is almost optimal. Thus, it should remain unchanged.
- AF exceeds $emax$: in order to prevent α from having a negative value, it should be reset to 0.5.
- AF remains positive for multiple successive adjustments: α could start to deviate to 0. If this is the case, α should be reset to 0.5.

The Algorithm 1 below illustrates the proposed method that is implemented on the sensor.

In contrast with other methods that require a set of data to be reported to the sink, in order to re-adjust the model, our approach requires only a single measurement to be transmitted. Thus, the adaptation phase is significantly faster, which reduces further radio communication. Moreover, the other approaches are far more complex and require

Algorithm 1: Transmission reduction.

```
Read $x_0$ and $x_1$
Transmit $x_0$ and $x_1$ to Sink
$d[0] \leftarrow x_1 - x_0$
$\alpha \leftarrow 0.5$
$L \leftarrow x_1$
$t_L = t_1$
while $Energy \neq 0$ do
    Read $x_l$ at time $t_l$
    $\widehat{x}_l \leftarrow \widehat{x}_{l-1} + d \times \alpha$
    if $|x_l - \widehat{x}_l| \geqslant emax$ then
        Send $x_l$ to Sink
        $d[l] \leftarrow \frac{x_l - L}{t_l - t_L}$
        $L \leftarrow x_l$
        $t_L \leftarrow t_l$
        $AF \leftarrow \frac{\widehat{x}_l - x_l}{N}$
        if $AF \geqslant emax$ then
            $\alpha \leftarrow 0.5$
        end if
        if $AF \leqslant \frac{10 \times emax}{100}$ then
            Do not update $\alpha$
        end if
        $P \leftarrow \frac{AF \times 100}{emax}$
        $\alpha^{new} \leftarrow \alpha - \frac{P \times \alpha}{100}$
        if $alpha \approx 0$ then
            $\alpha \leftarrow 0.5$
        end if
        $\widehat{x}_l \leftarrow x_l$
    else
        $N = N + 1;$
    end if
end while
```

a substantial amount of mathematical operations to readjust the model and to output a prediction.

1.3.2/ Identifying Wrong Predictions

Let us assume a scenario where a sensor fails to report a reading to the sink during the adaptation phase. The sensor will not know that the sink did not receive it, and it will use this reading to update its model. However, the sink considers that its prediction is within the error budget, therefore no update is needed. Hence, the prediction models on both sides will lose synchronization and start outputting different values. Therefore, we propose a solution that is based on an acknowledgment mechanism between the sensor and the sink. Consider that a sensor transmits a reading to the Sink, instead of switching immediately to the adaptation phase, the sensor must wait for an acknowledgment indicating that the reported value has been well received. As long as the sensor has not yet received an acknowledgment, it must keep reporting readings to the sink. This method ensures that both the sink and the node update their models simultaneously.

Moreover, a sequence number is sent with each reading. If the sink detects a jump in sequence numbers, it flags the corresponding measurements as missing, which allows the reconstruction algorithm to identify and reconstruct them.

Yet this is not sufficient to cover all potential failures, the batteries may deplete or the sensor could crash due to a software failure. Therefore, a sensor must operate in rounds, where each round is divided into several time slots. At the beginning of each time slot, a sensor must send the current reading even if it is within the error budget. Hence, if the sink does not receive a reading at the beginning of a given time slot, it will consider that the sensor has crashed, and all the future estimations will be replaced by a "NaN" value (flagged as missing).

The Algorithm 2 illustrates how the Sink can detect missing values.

Algorithm 2: Detecting missing data.

```
Receive $x_0$ and $x_1$
$d[0] \leftarrow x_1 - x_0$
$L \leftarrow x_1$
$t_L = t_1$
while $Energy \neq 0$ do
    if $x_l$ is received with sequence number $SN$ at time $t_l$ then
        Send an acknowledgment to the sensor
        if SN $>$ 1 then
            for j=1 to SN do
                $x_{l-j} \leftarrow$ "NaN"
            end for
        end if
        $d[t] \leftarrow \frac{x_t - L}{t_x - t_L}$
        $\widehat{x_t} \leftarrow x_t$
    else
        $\widehat{x_t} \leftarrow \widehat{x}_{t-1} + d \times \alpha$
    end if
end while
```

1.3.3/ Reconstruction of missing data

At the end of the sensing period, all missing data (NaN values) must be reconstructed. To achieve this, we have adopted and adapted the method proposed in [42]. This method exploits both temporal smoothness and spatial correlation among data sequences in order to estimate the values of missing measurements.

Let us consider a time sequence χ with duration T in m dimensions, where m is equal to the number of sensors in the monitoring area. This sequence χ contains all the data reproduced by the sink, and it also includes "NaN" values indicating that a reading is missing. The goal of this algorithm is to reconstruct these missing readings, by observing values of the missing sensor and other ones at neighboring time ticks.

$$\chi = \begin{bmatrix} x_1^1 & x_1^2 & ... & NaN & ... & x_1^T \\ x_2^1 & x_2^2 & ... & x_2^T & ... & x_2^T \\ ... & NaN & ... & & & \\ ... & & & & & \\ x_m^1 & x_m^2 & ... & NaN & ... & x_m^T \end{bmatrix}$$

Let us denote the observed part as χ_o, and the missing part as χ_m. A probabilistic model (Figure 1.1) is built to estimate the expectation of missing values conditioned by the observed part $\mathbb{E}[\chi_m|\chi_o]$.

A set of latent variables denoted Z_n are calculated using a belief propagation system. These latent variables model the dynamic and hidden patterns of the observed sequence. Moreover, the latent variables Z_n are assumed to be time-dependent with the value at time tick . It is determined by the value at time tick $t-1$ using a linear mapping F. In addition, linear projection matrix G from the latent variables Z_n to the data sequence for each time tick, is assumed to represent the spatial correlation among different dimensions.

Once the latent variables are calculated, they are used as input for an EM iterative algorithm [40] in order to find the best-fit parameters (such as G and F) for the data reconstruction probabilistic model.

Figure 1.1 illustrates this probabilistic model used to estimate missing values at a given time tick. For instance, the figure shows two missing values from two different sensor nodes at time tick 3. These values can be estimated using the linear projection matrix G and the estimated latent variable Z_3 at time tick 3. A detailed explanation of how the parameters of the model and the latent variables are calculated can be found in [40].

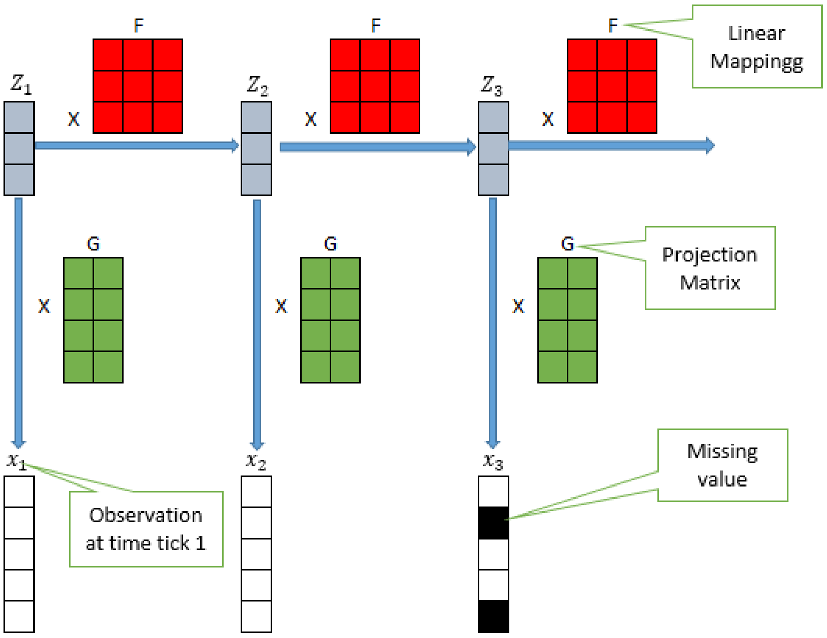


Figure 1.1: Probabilistic model

1.4/ Adaptive Sampling

In this section, we present our proposal for an adaptive model to calculate the sampling frequency of each sensor node based on the variance study. For this purpose, we use the One Way Anova model to determine whether there are any significant differences between the means of different data sets collected in successive periods. Then, different statistical tests (e.g. Fisher, Tukey, Bartlett, Kruskal-Wallis, etc.) can be used for comparing the factors of the total deviation.

1.4.1/ Study of measurements variance

In this part, we present a statistical model to calculate and analyze the means of measures taken by a node in order to adjust its sampling frequency. Within each period p, a sensor node takes several measures of temperature or humidity for example. To illustrate, we consider a sensor node after J periods. Our goal is to adapt the sampling rate in function of the new variance between periods, and this after every new period.

The variance between periods may be thought of as a signal of measures differences. Therefore, we use the one way ANOVA model to test if whether or not the means of several periods are all equal and if the variance differs from one period to another. We suppose that measures inside each period j are independent, with the mean $\overline{X_j}$ and the variances of periods are equal to σ^2. Then the variable of the $i-th$ measurement of period j, x_{ji} can be written as follows:

$$x_{ji} = \overline{X_j} + \epsilon_{ji}, \qquad j = 1, \ldots, J; \qquad i = 1, \ldots, n_j$$

where ϵ_{ji} are the residual which are independant and are normally distributed following $N(0;\sigma^2)$, and n_j is the number of the collected measures during period j.

We denote by:

$$\overline{X_j} = \frac{1}{nj}\sum_{i=1}^{nj} x_{ji}, \qquad \sigma_j^2 = \frac{1}{nj}\sum_{i=1}^{nj}(x_{ji} - \overline{X_j})^2, \qquad N = \sum_{j=1}^{J} n_j, \qquad \overline{X} = \frac{1}{N}\sum_{i=1}^{nj}\sum_{j=1}^{J} x_{ji}$$

the mean and the variance in each period and the mean of all the J periods respectively.

The total variation (ST) is the within period variation (SR) plus the between period variation (SF). The whole idea behind the analysis of variance is to compare the ratio of between period variance to within period variance. If the variance caused by the interaction between the measures is much larger when compared to the variance that appears within each period, then it is because the means are not the same. Let us consider:

$$ST = SR + SF \Rightarrow$$

$$\sum_{j=1}^{J}\sum_{i=1}^{nj}(x_{ji} - \overline{X})^2 = \sum_{j=1}^{J}\sum_{i=1}^{nj}(x_{ji} - \overline{X_j})^2 + \sum_{j=1}^{J} nj \times (\overline{X_j} - \overline{X})^2 \qquad (1.4)$$

1.4.1.1/ Data deviation verification

After studying the variance of data, now we use statistical tests to evaluate the changing of measures inter and intra periods. Different statistical tests can be used. For example, Fisher and kruskal-Wallis tests can be used as follows.

Fisher Test

Fisher test is a statistical hypothesis test for testing the equality of two variances by taking the ratio of the two variances and ensuring that this ratio does not exceed a certain theoretical value (that we can find in Fisher's table). Let:

$$F = \frac{SF/J-1}{SR/N-J} \tag{1.5}$$

If the hypothesis is correct then, F will have a Fisher distribution, with $F(J-1, N-J)$ degrees of freedom. The hypothesis is rejected if the F calculated from the measurements is greater than the critical value of the F distribution for some desired false-rejection probability (risk α). Let $F_t = F_{1-\alpha}(J-1, N-J)$. The decision is based on F and F_t:

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

Kruskal-Wallis Test

The Kruskal-Wallis H test is a rank-based non-parametric test that can be used to determine if there are statistically significant differences between two or more groups of data [43]. We propose to present the Kruskal-Wallis test via the following example.

Illustrative example Let us consider that a sensor operates in rounds, where each round R consists of p periods. To simplify the example, let us assume that R is equal to two. Table 1.1 shows a set of measurements collected by a sensor during two consecutive periods.

Table 1.1: Example of collected measures

Raw Measures		Measures Rankings	
Period 1	Period 2	Period 1	Period 2
3.4	4.6	1	2
6.2	5.8	4	3
7.0	7.0	~~5~~ 5.5	~~6~~ 5.5
7.3	7.5	7	8
7.6	8.0	9	10
10.3	10.2	~~12~~ 12.5	11
	10.3		~~13~~ 12.5
Number of Measures		Sum of Rankings	
6	7	39	52

The first step is to order the measurements in both periods by increasing order of their values and assign a rank denoted r to each one of them, representing its position in the ordered list. However, two or more measurements could have the same value. In this case the mean value of their ranks is calculated and assigned to each one of them. For instance, in Table 1.1 the value 7.0 is repeated twice, both in period 1 and 2 with ranks 5 and 6 respectively. The mean value of both ranks is 5.5. Thus, the ranks of both measurements holding the value 7.0 are replaced by 5.5.

The second step is to pass the ranked measurements as input to the Kruskal-Wallis test in order to find which one of the following assumptions is correct:

- Assumption 1: the two groups/sets of data (measurements in period 1 and 2) are significantly different.
- Assumption 2: the difference between the two groups of data is not significant.

The test is conducted by calculating the following formula :

$$H = \frac{12}{N \times (N+1)} \sum_{i=1}^{p} \frac{r_i^2}{n_i} - 3 \times (N+1) \tag{1.6}$$

where:

- N is the total number of measurements in all periods.
- n_i is the number of measurements inside the i^{th} period.
- r_i is the sum of all ranks in the i^{th} period.

Using the data in Table 1.1 and based on equation 1.6, H is calculated as follows:

$$H = \frac{12}{13 \times (13+1)}(\frac{39^2}{6} + \frac{52^2}{7}) - 3 \times (13+1) = 0.183$$

Finally, to check which assumption is the correct one, the result of this formula is compared with a "difference value" denoted H_t. H_t varies according to the false rejection probability predefined by the user, denoted α. The relation between α and F_t can be found in the $chi-square$ Table.

Let us assume α = 0.05, for this value of α, H_t is equal to 5.991. Comparing the results of the previous equation we notice that $H < H_t$ $(0.183 < 5.991)$. Therefore, the first assumption is accepted. Hence, the sampling rate must be adapted.

Remark 1: Statistical test notations

To simplify the notations, in the next sections we will use the notations F and F_t for the variation calculation and the test threshold. For example, for Kruskal Wallis test we only replace F by H and F_t by H_t.

After evaluating the variation of the collected data, the question now is "how to adapt the sampling rate according to this variation"?

1.4.2/ Adaptation To Application Criticality

A naïve approach to take into account the application criticality would consist in fixing the measure sampling rate of all sensor nodes to a given rate. As illustrated in Figure 1.2, we show how the sensor nodes capture speed can be regulated proportionally to the dynamic risk level r^0. For instance, a high criticality level pushes sensor nodes to capture at near the maximum sampling rate capability. The idea behind this model is, when the observed F becomes greater than the threshold F_t the sampling rate is balanced to the maximum sampling rate or to the minimum sampling rate in the other cases. However, this simple approach presents some drawbacks: (i) setting sensor nodes to work at full capacity provides high number of taken measures which need high bandwidth and leads to run out the sensor batteries and thus reducing the network lifetime, (ii) although setting the nodes at low capacity saves energy and extends the network lifetime, but it provides poor data quality where some important measures will be missed, (iii) choosing a moderate sampling rate could balance between capture quality and network lifetime but at the same time sensors cannot be fully exploited if it is necessary (when the physical changes are very dynamic).

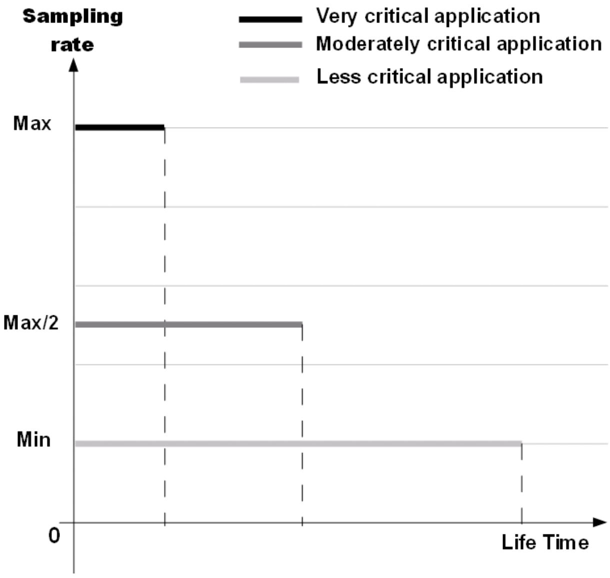


Figure 1.2: Naïve approach.

1.4.2.1/ Dynamic Sampling Model

To fully exploit the sensor node capabilities we propose that a node sampling rate depends on the result of the three tests as shown in Figure 1.3. Based on the results and residual of the variance test described above, the idea is that when a node noticed high variance differences, it can increase its sampling rate in order to prevent missing important measures and it decreases its sampling rate when the variance are lower than the threshold. In general, it is desirable to be able to adjust the sampling rate according to the application's requirements. In our approach we express the application criticality by the quantitative variable r^0 which can take values between 0 and 1 representing the low and the high criticality level respectively. Hence, taking into account the criticality value allows us to define in somehow an appropriate level of service. The higher the sampling rate is, the better relevant decisions and analysis could be made. Therefore a low criticality level indicates that the application does not require a high sampling rate while a high criticality level does. For instance, in health monitoring applications r^0 must be higher than snow monitoring applications. According to the requirements of the applications, an r^0 value that indicate the criticality level could be initialized accordingly into all sensors nodes prior to deployment and it can be adjusted online in function of the application requirement. Another example of using the r^0 value is in periodic healthcare applications. During monitoring phase of patient vital signs, we should consider the level of risk for each patient, i.e a patient considered in high risk situation at the hospital might be considered in high risk situation outside the hospital. Then, the adaptation algorithm must adapt its behavior by considering the risk level of the monitored patient. Patient with low level of criticality does not need continual monitoring, which means no need for high sampling rate, the case of patient with high risk level.

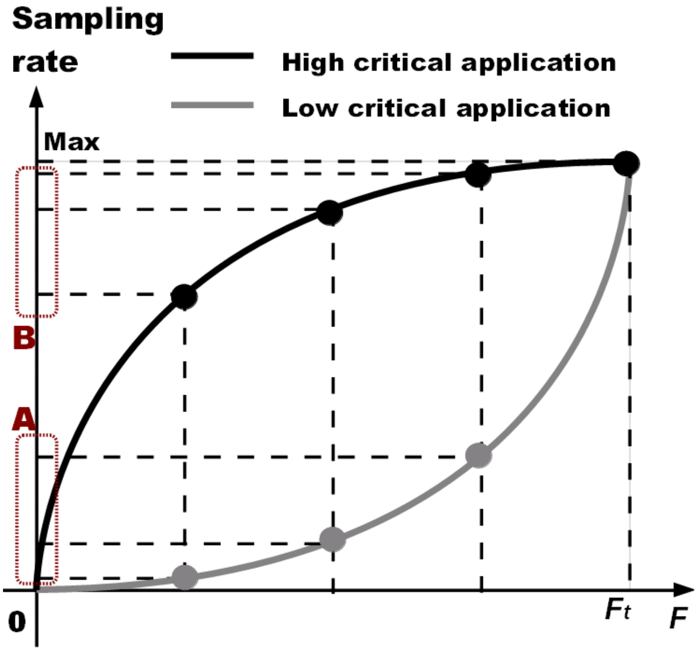


Figure 1.3: Dynamic approach.

1.4.2.2/ Influence on the F-ratio

In this section, we discuss why the sampling rate increases when the F-ratio increases. In fact, the greater the difference among the means, the higher the F and the greater the likelihood of obtaining variance differences. Hence, it is important to note that a large F does not by itself convey why or how the means differ from each other. A high F value can be found when the means for all of the groups differ at least moderately from each other. Alternatively, a high F can be obtained when most of the means are fairly similar but one of the means happen to be far removed from the other means. In this last case, the F ratio is influenced by group means, where the variances intra groups are very different. In these two cases, the sensor node must increase its sampling rate to capture all the physical changes.

1.4.2.3/ Application classes

We can broadly classify applications into different categories based on their criticality level. In our approach we define two classes of applications: high and low criticality applications. This criticality level is represented by a mathematical function $y = f_{r^0}(F)$ that we call BV (**B**eha**V**ior) function. This function can oscillate from hyperbolic to parabolic shape as illustrated in Figure 1.3:

- values on the x axis are positive results of the test (Fisher, Tukey or Bartlett). These values range between 1 and F_t. We consider that F_t corresponds to the maximum sampling rate. Thus, if a node finds a F greater than F_t it puts its sampling rate to the maximum.
- the y axis gives the corresponding sampling rate based on the test results on the x axis and the application criticality level (r^0) (number of sensed measures per unit of time).

We now present the contrast between applications that exhibit high and low criticality level in terms of the BV function.

1. **Class 1 ”low criticality”, $0 \leqslant r^0 < 0.5$:** this class of applications does not need high sampling rate. This characteristic is represented by an hyperbolic BV function. As illustrated in Figure 1.3 (box A), most projections of x values are gathered close to zero (i.e. the majority of the sensors will preserve their energy by sampling slowly).
2. **Class 2 ”high criticality”, $0.5 \leqslant r^0 \leqslant 1$:** this class of applications needs high sampling rate. This characteristic is represented by a parabolic BV function. As illustrated in Figure 1.3 (box B), most projections of x values are gathered close to the max frame capture rate (i.e. the majority of nodes capture at a high rate).

1.4.2.4/ The Behavior function

We use Bezier curve to model the BV function. Bezier curves are flexible and can plot easily a wide range of geometric curves.

Definition 1: bezier curve

The bezier curve is a parametric form to draw a smooth curve. It is fulfilled through some points $P_0, P_1 ... P_n$, starting at P_0 going towards $P_1 ... P_{n-1}$ and terminating at P_n.

In our model we will use a Bezier curve with three points which is called a Quadratic Bezier curve. It is defined as follows:

Definition 2: B(t)

A quadratic Bezier curve is the path traced by the function B(t), given points $P_0, P_1,$ and P_2.

$$B(t) = (1-t)^2 \times P_0 + 2t(1-t) \times P_1 + t^2 \times P_2. \tag{1.7}$$

The BV function is expressed by a Bezier curve that passes through three points:

- The origin point ($P_0(0,0)$).
- The behavior point ($P_1(b_x,b_y)$)
- The threshold point ($P_2(h_x,h_y)$) where h_x represents the highest cover cardinality and h_y represents the maximum frame capture rate determined by the sensor node hardware capabilities.

As illustrated in Figure 1.4, by moving the behavior point P_1 inside the rectangle defined by P_0 and P_2, we are able to adjust the curvature of the Bezier curve. The BV function describes the application criticality. It takes $|Co|$ as input on the x axis and returns the corresponding ”frame capture rate” on the y axis. To apply the BV function with the Bezier curve, we modify this latter to obtain y as a function of x, instead of taking a temporal variable t as input to compute x and y. Based on the Bezier curve, let us now define the ”BV function”:

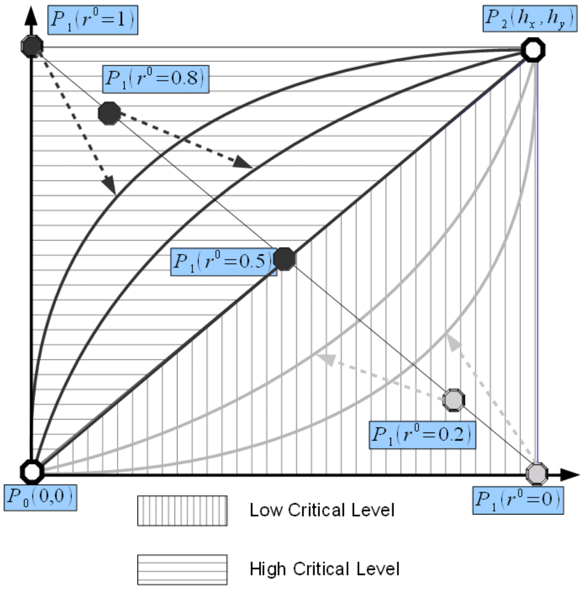


Figure 1.4: The Behavior curve functions.

Definition 3: BV function

The BV function curve can be drawn through the three points $P_0(0,0)$, $P_1(b_x,b_y)$ and $P_2(h_x,h_y)$ using the Bezier curve as follows:

$$\begin{aligned} BV : [0,h_x] &\longrightarrow [0,h_y] \\ X &\longrightarrow Y \end{aligned}$$

$$BV_{P_1,P_2}(X) = \begin{cases} \frac{(h_y-2b_y)}{4b_x^2}X^2 + \frac{b_y}{b_x}X & if\ (h_x - 2b_x = 0) \\ (h_y - 2b_y)(\propto(X))^2 + 2b_y\propto(X), & if\ (h_x - 2b_x \neq 0) \end{cases} \qquad (1.8)$$

$$Where\ \propto(X) = \frac{-b_x+\sqrt{{b_x}^2-2b_x\times X+h_x\times X}}{h_x-2b_x} \wedge \begin{cases} 0 \leqslant b_x \leqslant h_x \\ 0 \leqslant X \leqslant h_x \\ h_x > 0 \end{cases}$$

1.4.2.5/ The criticality level r^0

As discussed above, the criticality level r^0 of an application is given into the interval $[0,1]$. According to this level, we define the criticality function called Cr which operates on the behavior point P_1 to control the BV function curvature.

According to the position of point P_1 the Bezier curve will morph between parabolic and hyperbolic form. As illustrated in Figure 1.4 the first and the last points delimit the curve frame. This frame is a rectangle and is defined by the source point $P_0(0,0)$ and the threshold point $P_2(h_x,h_y)$. The middle point $P_1(b_x,b_y)$ controls the application criticality. We assume that this point can move through the second diagonal of the defined rectangle $b_x = \frac{-h_y}{h_x} \times b_y + h_y$.

We define the Cr function as follows, such that varying r^0 between 0 and 1 gives updated positions for P_1:

$$\begin{array}{rcl} Cr : [0,1] & \longrightarrow & [0,h_x] \times [0,h_y] \\ r^0 & \longrightarrow & (b_x,b_y) \end{array}$$
$$Cr(r^0) = \begin{cases} b_x = -h_x \times r^0 + h_x \\ b_y = h_y \times r^0 \end{cases} \tag{1.9}$$

Level r^0 is represented by the position of point P_1. If $r^0 = 0$ P_1 will have the coordinate $(h_x,0)$. If $r^0 = 1$ P_1 will have the coordinate $(0,h_y)$.

1.4.3/ Adapting sampling rate Algorithm

In this section, we present the adaptive sampling rate algorithm.

Algorithm 3 describes the adaptive sampling rate algorithm at the sensor node level. For each round, every node decides to increase or decrease its sampling rate according to the difference between its collected measures and the application risk level. While the energy is always positive, each node calculates the parameters F, F_t, according to the used statistical test. Then, it uses the Behavior function to adapt its sensing rate only if the calculated difference between measures is less than the test threshold.

Algorithm 3: Adaptive Sensing Rate Algorithm.

```
Require: $R$ ($R$ = $p$ periods), $\tau$: period size, $r^0$: application criticality, $\alpha$: false-rejection probability.
Ensure: $S_t$ (instantaneous sampling speed).
$S_t \leftarrow \tau\ measures/period$
while $Energy > 0$ do
    for $i = 1 \rightarrow p$ do
        takes measures at $S_t$ speed
    end for
    for each round do
        sort measures by increasing order of their values. (Only for Kruskal-Wallis test)
        compute the rank of each measure (Only for Kruskal-Wallis test)
        find $F_t$
        if $F \leqslant F_t$ then
            $S_t \leftarrow BV(F, F_t, r^0, \tau)$ (BV behavior function).
        else
            $S_t \leftarrow \tau\ measures/period\ (S_max)$
        end if
    end for
end while
```

1.5/ Dual Prediction and Adaptive Sampling Approach

The Adaptive Sampling algorithm (AS) reduces the sampling rate of a sensor when the difference between collected measurements is not significant. Thus, enabling the sensor node to avoid collecting redundant and superfluous information. The Transmission Reduction algorithm (TR) reduces the number of data transmitted to the Sink, using a prediction model that can forecast future measurements within a narrow error range. The efficiency of the prediction model is at peak when data is smoothly changing with low variance between measurements.

Thus, one can notice that these two techniques are complementary to each other. The prediction model is capable of filling the gap of ”non collected data”, since as mentioned before these measurements are mostly redundant or roughly similar to closely collected ones, and the prediction model efficiency is at maximum when the change in values is smooth and slow. Therefore, on the one hand the sampling rate is reduced and on the other hand, the end user will still have access to the complete set of data. Finally the complexity of the transmission reduction algorithm is extremely low. Thus, when combined with the adaptive sampling algorithm the overall complexity will remain unchanged. Therefore, we propose to combine these two techniques into a single algorithm, enabling us to achieve lower energy consumption compared to each one of them when implemented solely.

Let us begin by explaining how to combine these two techniques together, and how the algorithm works as a whole. As mentioned earlier, the operations conducted by AS and TR does not overlap, and they do not affect the results of each others. Therefore, since they are totally compatible they can be implemented as they are originally without any changes in the way of working.

The only difference is, instead of having a single sampling rate, the sensor has two: a real one and a hypothetical one. The real sampling rate is the rate defined by the AS algorithm after each round. The hypothetical one is a fixed rate that is always equal to the maximum sampling rate. The sensor collects measurements at the real rate speed returned by AS. However, it uses the hypothetical rate while applying the TR algorithm. In other words, let us suppose for a round R, the real rate is RR measures/period and the hypothetical rate is HR measures/period. In this case, during this period, on the one hand the sensor collects RR measurements, on the other hand it predicts HR measurements. Note that RR is always less than HR, as HR is equal to the maximum sampling rate allowed for the sensor. Thus, the sensor is able to predict the ”non-collected” measurements caused by a slowed down sampling rate forced by AS.

Figure 1.5 gives an illustrative example on how this algorithm behaves.

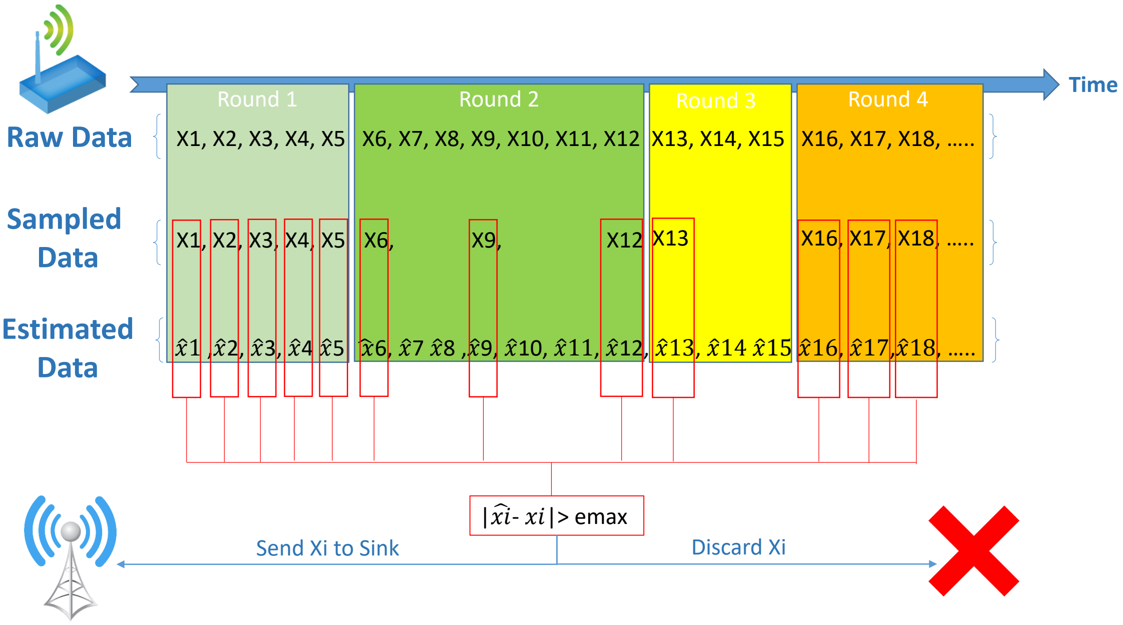


Figure 1.5: Data Prediction and Adaptive Sampling mechanism

We have explained in Section 1.3 that every time a sensor predicts a new measurement it must compare it with the real sensed value in order to decide whether it should send it to the Sink or not. However, when adding AS to the equation, if $RR \leqslant HR$ some predictions might not have a matching sensed measurements to validate its accuracy. Therefore, these predictions are considered to be within the error budget automatically. Since the ”non-collected” measurements are assumed to be redundant or already similar to collected neighboring measures. This assumption should not affect the accuracy of replicated data. However, we will show and discuss this issue in the Experimental Results section below.

1.6/ Performance Evaluation

In this section, we present the experimentation we have conducted on real wireless sensor readings, collected by twenty Crossbow TesloB nodes deployed in our laboratory as shown in Figure 1.6. The walls separating the rooms of the lab are of a thickness of 8 cm. Each one of the twenty nodes collects five environmental features: temperature, humidity, light, infrared and voltage. The collected measurements are directly transmitted to a central node (Sink) called SG1000, connected to a laptop machine, through a star topology.

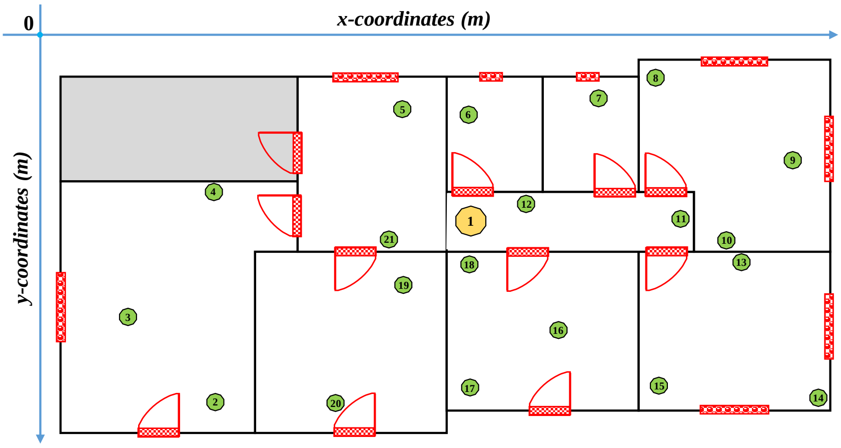


Figure 1.6: Deployment of the sensor network in our laboratory

We have chosen to test our approach on temperature, humidity, and infrared data since the variation in measured values for each one of them is different. For instance, the variation in temperature data is low, medium for humidity, and high for infrared. Thus, we can assess the efficiency of our approach on different scenarios.

1.6.1/ Dual prediction performance

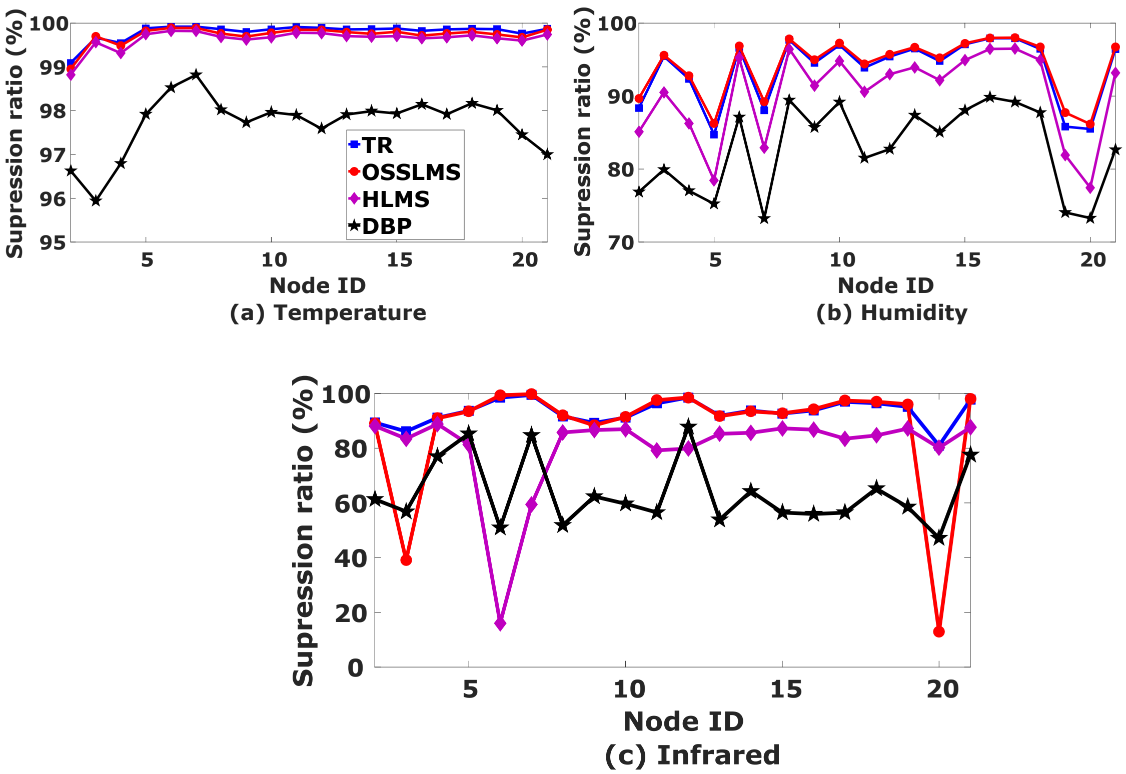
The performance evaluation was conducted on twenty sets of 300,000 readings each (100,000 temperature, 100,000 humidity, and 100,000 infrared readings), which is equivalent to approximately 35 days of non-stop data collection. For instance, temperature data tend to be very smooth, which makes it easy to predict future readings. Humidity data tend to vary faster and more frequently than temperature, therefore, it is harder for the model to keep up with the changes. Finally, infrared data does not follow any specific upward or downward trend like humidity and temperature do. The variations in collected measurements are either steady, or irregular and intense, which puts the prediction model in a constant adjustment state. Furthermore, To evaluate our approach (TR), we consider the available data sets described earlier and compare the effectiveness of our transmission reduction method to three state of the art linear data reduction algorithms (OSSLMS [22], HLMS [21], and DBP [23]).

Remark 2: Complexity comparison

We notice that our approach (TR) has the lowest complexity ($O(1)$) comparing to OSSLMS ($O(n^3)$), HLMS ($O(n^2)$) and DBP ($O(n)$). It seems that the higher the complexity of the algorithm, the better its efficiency. However, our method is the least complex one and achieves the best results.

1.6.1.1/ Data Transmission reduction

In this experimentation we fixed the error threshold (emax) to ± 0.1 for temperature and humidity, and ± 1 for Infrared. For the HLMS algorithm, the number of sub-filters (m2) is set to 2 and the size of each one of them (m1) to 3. We have selected the most optimal step size parameter μ $(3 \times 10^{-4}$ for temperature, 2×10^{-6} for humidity and infrared), in order to compare our method with the most efficient HLMS. For the OSSLMS algorithm the size of the filter was set to 5 and for the DBP algorithm the number of edge points l was set to 3, the learning window m fixed to 6, the relative error 5% for all environmental features, and the time tolerance ϵ_T to 2. We evaluated the suppression ratio (SR) which is the percentage of collected data and not sent to the sink. The final averaged results are listed in Table 1.2. The ratio of suppressed transmissions of each individual node for all of the twenty data sets are shown in Figure 1.7.



(a) Temperature

(b) Humidity

(c) Infrared

Figure 1.7: Suppression ratio for each individual node

All the proposed algorithms performed well in terms of data suppression. However, our data reduction method outperformed the other approaches in two out of three environmental features, and OSSLMS has outperformed our method in one feature.

Table 1.2: Suppression ratio of transmitted data

	Supression ratio (%)			
	TR	OSSLMS	HLMS	DBP
Temperature	**99.8**	99.7	99.6	97.7
Humidity	93.6	**94.1**	90.3	82.7
Infrared	**93.2**	87.6	80.1	63.5

As shown in Table 1.2, all of the four algorithms have approximately the same Suppression Ratio (SR) for temperature data, except for DBP that has an SR that is around 2% lower than the others. The reason is that temperature data are very smooth and the variations in neighboring measured values are small. Thus, a linear prediction algorithm is very efficient in keeping up with these small changes. For humidity data, OSSLMS has adapted itself better and achieved the best suppression ratio of 94.1% among the three other approaches, including ours that achieved an SR of 93.6%. Finally, when we tested the algorithms on highly varying infrared data, our approach was significantly better in reducing the number of radio communication. For instance, our SR was 5.6%,13.1%, and 29.7% greater than OSSLMS, HLMS, and DBP respectively.

1.6.1.2/ Data loss/Communication error

In this section, we will compare the different approaches in a scenario where data loss occurs randomly when a sensor is trying to send a measurement to the Sink. The percentage of a sensor failing to transmit its readings can vary from 10% in an ideal environment and when there is low collision and overload in the network, up to 50% in harsh environments and a jammed network [44]. We have evaluated the performance of each of the four methods with a transmission failure possibility ranging from 10% to 50% based on the Bernoulli distribution, where the probability of failed transmission p is varied within the range of $[0.1 - 0.5]$, and the probability of a successful transmission q $(1 - p)$ within $[0.5 - 0.9]$.

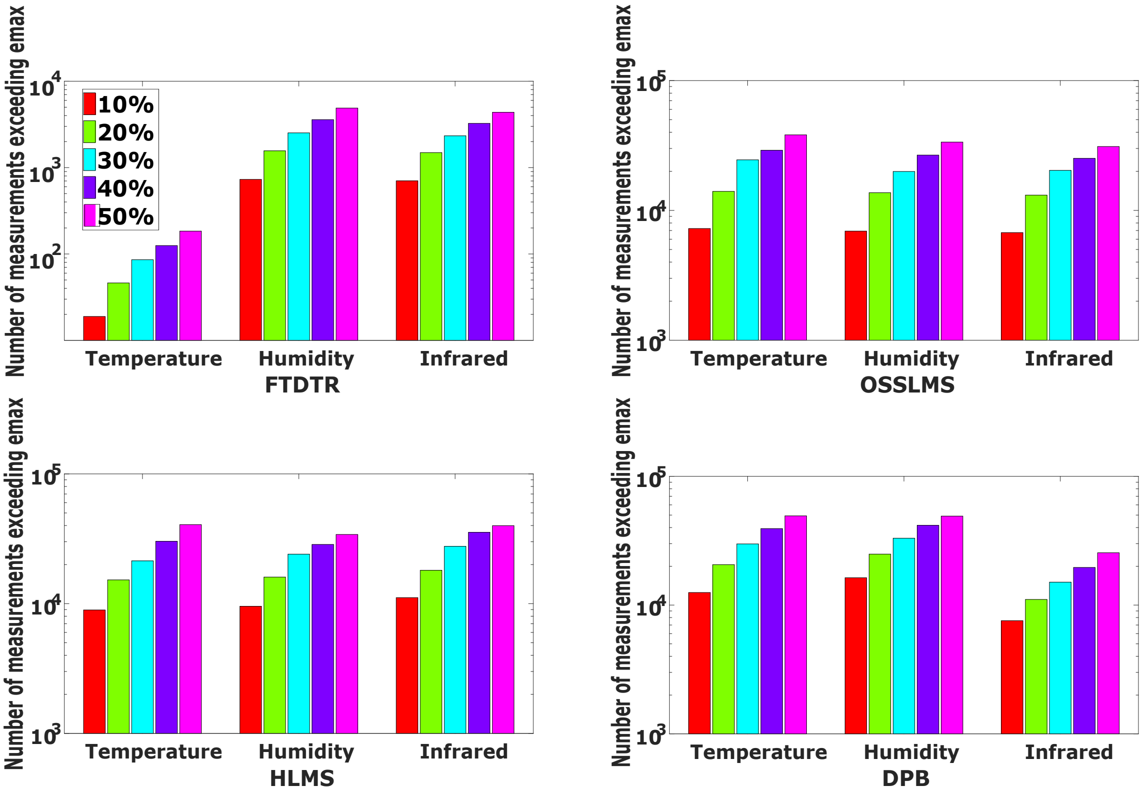


Figure 1.8: Amount of data exceeding "*emax*"

Figure 1.8 shows the number of temperature, humidity, and infrared data exceeding the error threshold $emax$ when different missing possibilities are considered. With a data loss detection mechanism, our approach ”fault tolerant data transmission reduction” (FTDTR) was able to limit the number of wrong estimations by keeping the model at the sensor and the sink synchronized. Thus, only the readings that failed to reach the sink and were flagged as ”NaN” values by the latter are considered to exceed the error threshold. Oppositely, the number of estimations exceeding the error threshold for other approaches is far greater than the number of measurements that failed to be reported. The reason is that the readings that fail to reach the sink are used by the sensor to adjust the model, thus leaving the sink with an outdated one that produces wrong estimations, while the sensor is producing correct ones.

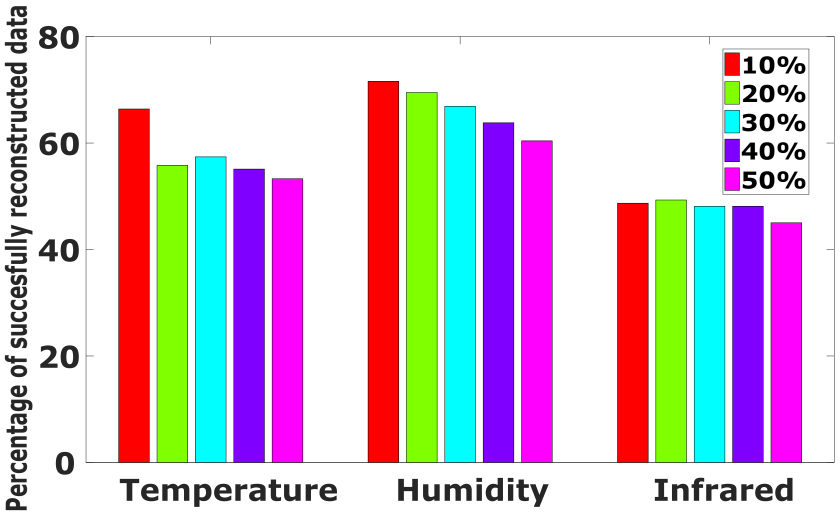


Figure 1.9: Percentage of successfully reconstructed data

For each environmental feature (temperature, humidity, and infrared) the twenty data sets reproduced by the sink corresponding to the twenty deployed sensor nodes are passed to the reconstruction algorithm in order to fill the blank values flagged as ”NaN”. The number of latent variables was set to 20 (number of time series data to be reconstructed).

Figure 1.9 shows the average percentage of the successfully reconstructed data. The reconstruction success rate can range between 45% and 72% according to the number of missing readings and on the temporal smoothness and spatial correlation of the data set. The reconstruction of a missing reading at time t is considered to be unsuccessful if the difference between the values of the reconstructed measurement and the real one is greater than the maximum error tolerance ($|\hat{x}_t - x_t| > emax$).

For a missing probability varying from 10% to 50%, Table 1.3 shows the Root Mean Square Error (RMSE) of:

- the measurements that have been unsuccessfully reconstructed by FTDTR,
- the data exceeding $emax$ for OSSLMS, HLMS, and DBP.

The results show that our method has the lowest RMSE for all environmental features and for all missing probabilities. Moreover, for temperature and humidity data, the RMSE’s are very close to $emax$ (0.1). Therefore, when we increased $emax$ to 0.2, the reconstruction

Table 1.3: RMSE of data exceeding emax

	Temperature				Humidity				Infrared			
	FTDTR	OSSLMS	HLMS	DBP	FTDTR	OSSLMS	HLMS	DBP	FTDTR	OSSLMS	HLMS	DBP
10%	0.103	0.16	0.22	0.639	0.13	0.41	1.88	0.44	2.3	3.21	2.84	10.65
20%	0.104	0.188	0.238	0.849	0.15	0.29	2.15	0.47	2.7	3.39	4.39	15.52
30%	0.104	0.218	0.356	1.233	0.14	0.28	2.424	0.763	2.8	3.45	4.85	15.28
40%	0.111	0.242	0.782	1.025	0.14	0.312	2.937	0.834	2.8	4.06	5.42	18.46
50%	0.110	0.252	3.04	1.10	0.16	0.472	3.374	0.838	3.3	4.41	6.82	22.31

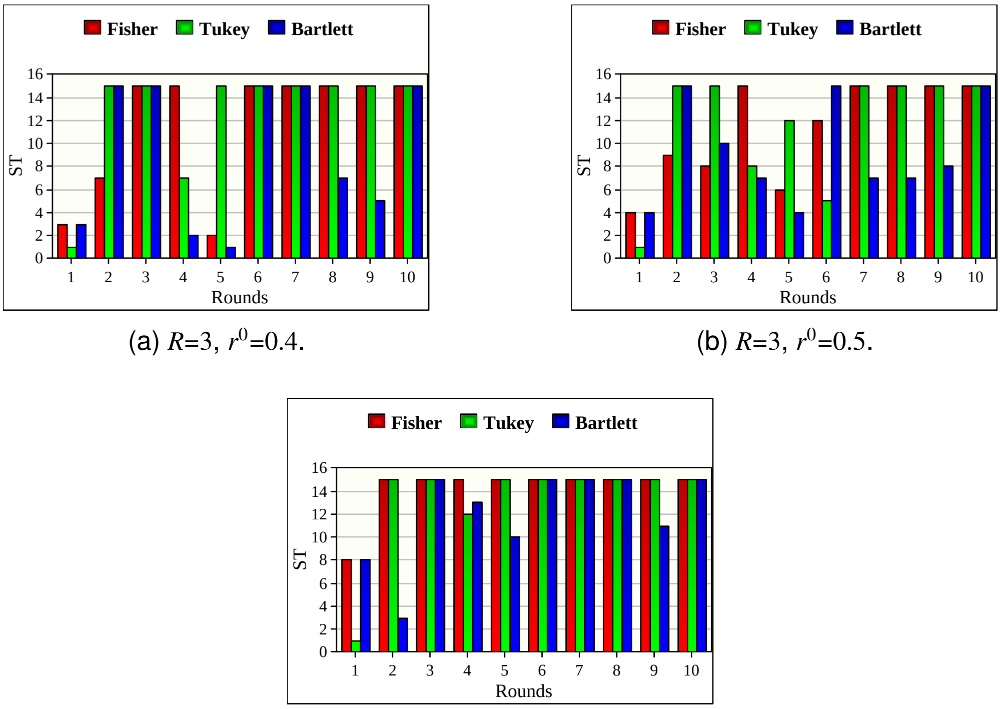
success rate for a 50% miss probability reached 99.6% and 94.3% for temperature and humidity data respectively. For infrared data when we increased $emax$ to 4 instead of 1 (which is still an acceptable error for most applications) the reconstruction success rate increased to 91%.

The obtained results are in line with what has been previously emphasized, showing that the proposal outperforms existing works in maintaining high-quality estimations when a data loss scenario occurs.

1.6.2/ Adaptive Sampling Performance

1.6.2.1/ Instantaneous Sampling Rate and application criticality

The main goal of this section is to show, on the one hand, how our approach is able to reduce and to adapt its sampling rate according to the application criticality level, and in other hand, to compare the results of the three different statistical tests Fisher, Tukey and Bartlett.



(a) R=3, r^0=0.4.

(b) R=3, r^0=0.5.

(c) R=3, r^0=0.8.

Figure 1.10: Variation of sampling rate (ST) over rounds, $SMAX = 15$, $\alpha = 0.05$.

Figure 1.10 shows the instantaneous sampling rate results for the three tests. In figures a, b and c we fixed each round to three periods ($R = 3$) when we varied the criticality level (r^0) to 0.4, 0.5 and 0.8 respectively. The period is fixed to 5 minutes, the maximum sampling rate $SMAX$ to 15 measures/period, and $\alpha = 0.05$. Based on these figures, we can see that the three tests successfully adapt the sampling rate of the sensor nodes dynamically after each round according to the application criticality level. These results show how the sampling rate ST varies over time. They confirm also the reduction of the amount of collected data comparing to the nodes operating on $SMAX$ all time. We can also observe that when the risk increases the sampling rate remains usually at its maximum value.

1.6.2.2/ Data size reduction

To show the effectiveness of our approach in reducing the size of data sent to the sink, in this experimentation we compare Kruskal Wallis test, Bartlett test and a data compression technique S-LEC [45].

Table 1.4 shows the comparison for the two fields, temperature and humidity. Every 30 seconds, each mote collects a new measure of each field. The application criticality r^0=0.6, and the false-rejection probability $\alpha = 0.01$.

Table 1.4: Data reduction ratio

Variable	Bartlett	Kruskal-Wallis	S-LEC
Temperature	81%	79.6%	74%
Humidity	74%	71.5%	66%

The obtained results show that Kruskal-Wallis and Bartlett tests outperforms the S-LEC techniques in terms of data transmission reduction for both temperature and humidity.

Remark 3: Statistical tests

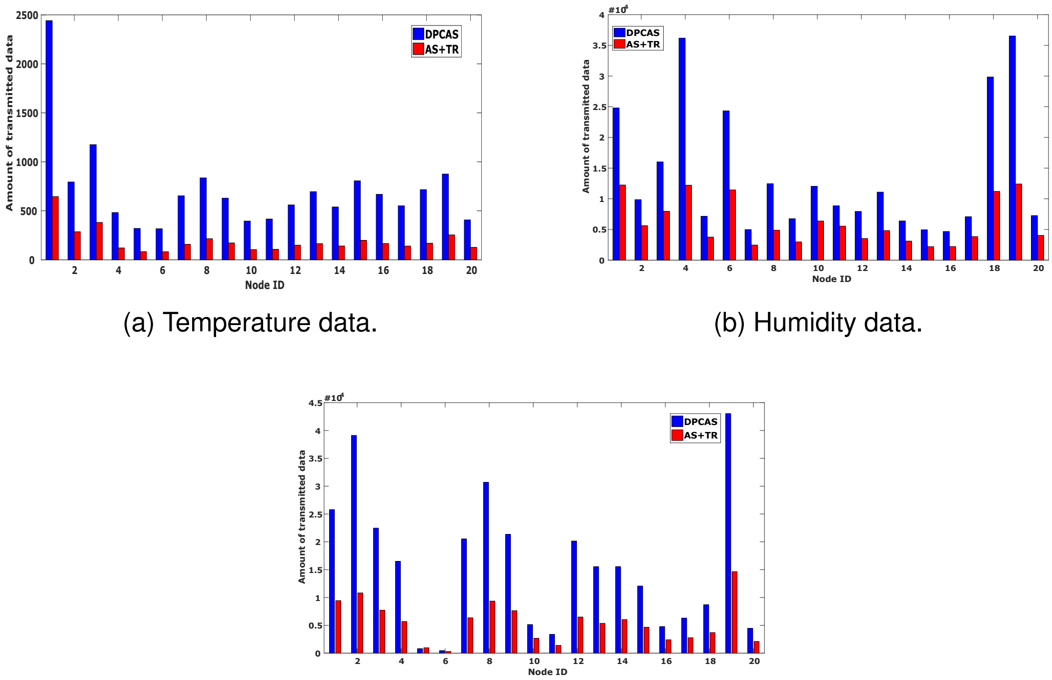
The objective of a statistical test is to verify if the variation between sets of data is above a certain threshold. In our work, we compared four different tests, Fisher, Tukey, Bartlett and Kruskal-Wallis [8, 9]. Bartlett's Test seems to be the most uniformly powerful test for the homogeneity of variances problem in the case that the data are normal. However, it has a serious weakness if the normality assumption is not met. Consequently, in such case, we must adopt other tests like Kruskal-Wallis which is the best test in reducing data loss.

1.6.3/ Data Prediction and Adaptive sampling performance

In this section we show the results obtained while combining the two data collection methods (Adaptive sampling and dual prediction Transmission reduction (AS+TR)). Furthermore, we compared the results to a recent published approach DPCAS [41]. The application Risk r^0 was fixed to 0.6 and the false rejection probability α to 0.05. We used the same S_{max} and S_{min} for DPCAS, the smoothing coefficient α and the multiplicative reduction factor β were fixed to 0.2, and the cubic parameter C to 0.4. The error threshold ”emax” was set to $\pm$0.1 for temperature and humidity, and $\pm$1 for infrared.

1.6.3.1/ Data Transmission reduction

Figures 1.11a, 1.11b and 1.11c show a comparison between the amount of temperature, humidity, and infrared data that have been transmitted to the sink when both DPCAS and AS+TR were implemented. For instance, on average, only 193 temperature readings or 0.193% were transmitted when our algorithm was implemented and 714 (0.714%) for DPCAS. For humidity and infrared, the results were identical, our method outperformed DPCAS in reducing the number transmissions. The average amount of transmitted humidity and infrared data for AS+TR is 6148 and 5528 respectively. As for DPCAS the numbers increase to 13964 and 15848 respectively. Hence, these results show that our transmission reduction method is better at reducing radio communication, which enables the node to preserve more of its energy resources.



(a) Temperature data.

(b) Humidity data.

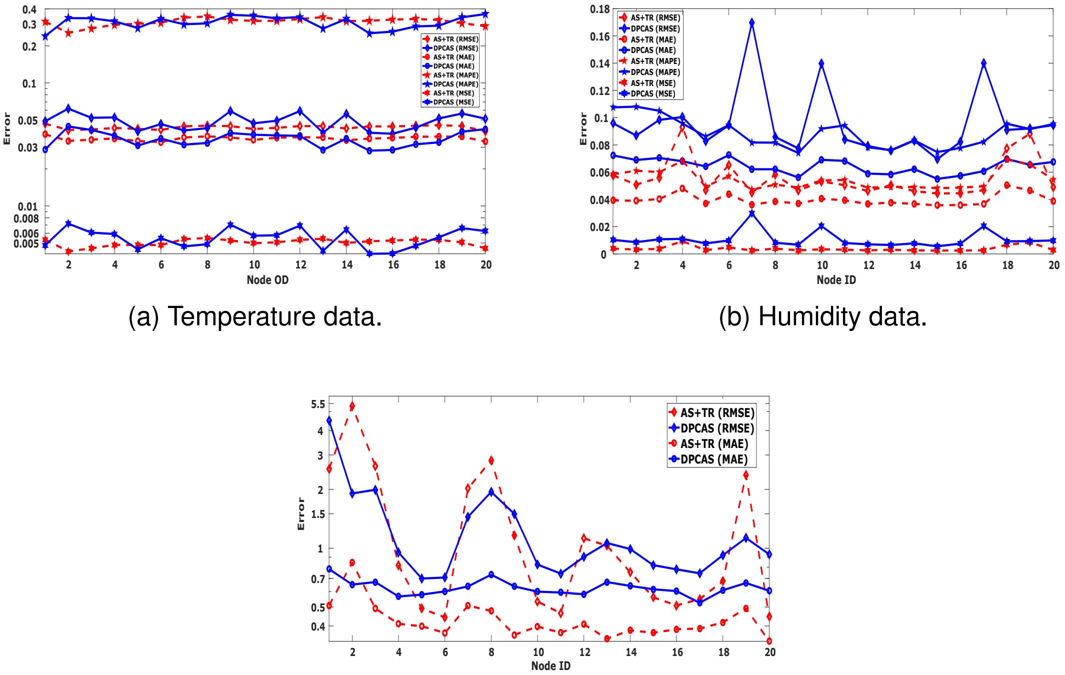
(c) Infrared data.

Figure 1.11: Amount of data transmitted to the Sink.

1.6.3.2/ Quality of replicated data

Data quality is a very important factor in WSNs, since the end user depends on it to make appropriate decisions. Accuracy, precision, completeness, and consistency are the attributes that measure the quality of data. When we reduce the sampling rate within a certain period, we risk missing sudden variations in measurements. Thus, the estimation of these irregular non sampled data may exceed the desired error threshold.

To study the impact of the adaptive sampling algorithms on the integrity of the replicated data we compare the estimated measurements with its corresponding raw data collected by the sensor node, and we calculate the values of 4 quality metrics: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE), Root Mean Square Error (RMSE). The lower the values of these metrics the better are the results.



(a) Temperature data.

(b) Humidity data.

(c) Infrared data.

Figure 1.12: Quality metrics comparison for replicated data.

Figures 1.12a, 1.12b, and 1.12c show a comparison between the quality metrics for each set of data of the three environmental features. For temperature data, both algorithms are neck to neck. However, AS+TR has a clear superiority for humidity and Infrared. Since Infrared contains 0 elements the value of MPAE cannot be computed. As for MSE in order to keep Figure 1.12c simple and comprehensible, instead of plotting the curve we provide the average values which are 2.49 and 2.72 for AS+TR and DPCAS respectively. Adaptive sampling makes a trade-off between data quality and the amount of sampled measurements, to deliver a minimum amount of readings while satisfying quality requirements of the application. Thus, the integrity of data depends on how tolerant is the end user to the error in replications. The obtained results demonstrated that our method was able to reproduce the whole data set with less error and better quality compared with DPCAS.

1.7/ Conclusion

In this chapter, we studied data collection in WSN. We proposed an energy-efficient data reduction method for Wireless Sensor Networks based on a combination of the adaptive sampling and dual prediction mechanism techniques. The former, allows the sensor to adapt its sampling rate according to the environment condition changing. Thus, the sensor samples relevant data only and avoids the sampling of redundant and insignificant information. The latter enables the Sink to estimate the collected data through a prediction mechanism that is shared with the sensor node. Thus, instead of transmitting all the readings, the sensor reports to the Sink a measurement only when the estimation exceeds a predefined error threshold. By merging these two techniques together, we were able to reduce radio communication and data sensing at the same time. Thus, preserving a great amount of energy and extending the network lifetime.

2

Data Aggregation in Large Scale WSN

Wireless sensor networks (WSNs) are being increasingly used in several applications and deployed in a densely distributed manner. Huge amount of data are usually sensed in these large-scale networks, which greatly affect the life of sensor nodes. Hence, in-network data aggregation is considered an effective technique to reduce the massive volumes of data by eliminating the redundancy and thus conserving energy communication in WSN. In this chapter, we propose and compare three different data aggregation techniques for cluster-based WSN. Further to a local aggregation at the sensor node level, our proposal allows each cluster-head (CH) to eliminate redundant data sets generated by neighboring nodes. The three proposed methods are based respectively on sets similarity functions, K-means algorithm and one-way Anova model, and distance functions. Based on real sensor data, we have analyzed their performances according to the energy consumption and the data latency and accuracy, and we show how these methods can significantly improve the performance of the network.

2.1/ Introduction

Wireless sensor networks are usually deployed to collect environmental data from a region of interest. We consider the scenario in which a large number of sensor nodes are densely deployed and tremendous amount of sensory data are measured and forwarded to the sink on a periodic basis. In this scenario, the collected data are spatially correlated where neighboring nodes operate with identical or similar sampling rates and redundant packets are likely to happen repeatedly. Therefore, it is important to take full advantage of the correlations among the sensor measures to reduce the size of the transmitted data and thus the cost of communication and energy consumption. In-network data aggregation has been proven as an effective technique for eliminating redundancy and forwarding only the extracted information from the raw data, resulting in conservation of energy and bandwidth.

In this chapter, we consider that each sensor node monitors the given phenomenon and periodically sends its collected data to its cluster head (CH). Then, we present a two layers data aggregation scheme aiming to optimize the volume of data transmitted thus saving energy consumption and reducing bandwidth requirements of the network. A first layer in-sensor process is done by the nodes themselves. Instead of sending raw data, each sensor node reduces redundancies from the collected data before transmitting it to the CH for a second layer of aggregation. We provide a non-complex algorithm based on a distance function we called it $Similar$ between readings. At the level of CH, we are interested in exploring a new part of the aggregation problem, by focusing on identifying the similarity between data sets generated by neighboring nodes. Our objective is to identify similarities between near sensor nodes, and integrate their captured data into one record while preserving information integrity. At this level, we present and compare three different methods:

- The first method is based on the sets similarity functions (e.g. Jaccard function) to search the similarities between data sets. A new optimization method for early termination of sets similarity computing is proposed.
- The second method uses an enhanced k-means clustering method with one-way ANOVA model and statistical tests in order to identify neighboring nodes generating close data sets.
- The third technique exploits the distance based functions (e.g. Euclidean and Cosine) to find near data sets with the aim to eliminate redundancies and reduce the huge amount of data transmitted over the network.

We studied the performance of our proposed techniques via simulations and real sensor networks implementation. The obtained results show the effectiveness of these techniques on reducing the cost of communication and energy consumption. Then, we compared the three proposed techniques in order to let the user choose the best solution that matches the most with the application requirements.

The rest of this chapter is organized as follows: Section 2.2 describes some of the existing data aggregation techniques proposed for WSNs. Section 2.3 provides a brief discussion of the data aggregation system based on clustering topology. In Section 2.4, we describe the aggregation at the sensor level, called local aggregation. Sections 2.5, 2.6 and 2.7 present the three aggregation methods at the CH level based on the similarity functions, the variance study and the distance functions respectively. Section 2.8 details the experimentation we have conducted in real data with some discussions. Finally, Section 2.9 concludes this chapter.

2.2/ State of the art

Reducing energy consumption is a major issue in WSNs where sensors are resource constrained. Hence, data aggregation is an essential operation in WSNs used to decrease the data transmission, thus, to enhance the network lifetime. It means computing and transmitting partially aggregated data to the end user rather than transmitting raw data. Data aggregation in wireless sensor networks has been well studied in recent years [46, 47, 48, 49, 50]. Indeed, these techniques are based on the network's topology. In the literature one can find data aggregation techniques for different topology, such as Tree-based [51], Cluster-based [52], Chain-based [53] or structure free-based [47] topology.

The authors in [46, 54, 55], use the clustering methods for aggregating data packets in each cluster separately. In [56], the authors propose a data aggregation scheme named

DMLDA, Dynamical Message List based Data Aggregation, based on clustering routing algorithm. DMLDA mainly defines a special list structure to store history messages, which is used to judge the message redundancy instead of the period delay. In [57], the authors propose an aggregation and transmission protocol (ATP) based on clustering approach to conserve energy in PSNs. Instead of sending raw data to the CH, ATP allows each sensor node to eliminate redundancy among its collected data and to adapt its data transmission to the CH, while studying the variance of the collected data.

Other proposed techniques of data aggregation are based on a tree network topology, such as [48, 49]. The authors in [48] use Genetic Algorithm (GA) to calculate all possible routes represented by the aggregation tree. The objective is to find the optimum tree which is able to balance the data load and the energy in the network. In [49], a semi-structured protocol based on the multi-objective tree is proposed, in order to reduce transmission delays and enhance the aggregation probability. In such a work, the routing scheme explores the optimal structure by using the Ant Colony Optimization (ACO).

Other data aggregation techniques are based on a chain routing topology [50, 58]. In [50], the authors propose a Cycle-Based Data Aggregation Scheme (CBDAS) in order to reduce the amount of data transmitted to the base station (BS). In CBDAS, the network is divided into a grid of cells, each with a head. The network lifetime is prolonged by linking all cell heads together to form a cyclic chain, where the gathered data move from node to node along the chain, getting aggregated. In [58], a chain-based routing scheme for application-oriented cylindrical networks is proposed. After finding local optimum paths in separate chains at each scheme, the authors formulate mathematical models to find a global optimum path for data transmission through their interconnection.

Finally, some works proposed recently are based on a structure-free network [59, 60]. In [59], the authors propose a Structure-Free and Energy-Balanced data aggregation protocol, SFEB. It features both efficient data gathering and balanced energy consumption, which result from its two-phase aggregation process and the dynamic aggregator selection mechanism. In [60], a virtual force-based dynamic routing algorithm (VFE) for data aggregation in WSNs is proposed. Motivated by the cost field and virtual force theories, VFE allows each node to select the optimal node to be the next hop which makes data aggregation more efficient.

Subsequently, clustering is recently considered as an efficient topology control method in WSN [42] that has many advantages, especially as far as scalability and network maintenance are concerned. However, most of the existing data aggregation techniques based on clustering topology are dedicated to event driven data model [61, 12] and they mainly focus on the selection of CHs [62, 63]. In these techniques only CHs process and aggregate data without any processing at the level of the nodes themselves. In this chapter, we present three data aggregation techniques dedicated to cluster based WSN. They have two layers which are able to eliminate redundant measurements and data sets generated by the nodes at each period.

2.3/ Cluster based Network topology

In this chapter, we focus on the cluster-based network topology where the whole network is divided into several clusters. Each cluster has a cluster-head (CH) which is responsible for managing the sensors in the cluster. Indeed, grouping sensor nodes into clusters has been widely studied by the research community to satisfy the scalability objective and achieve high energy efficiency and prolong the network lifetime [64, 65, 66]. Some of the proposed techniques aim at forming and maintaining the clustered networks while optimizing the cluster size [67, 68], others try to elect smartly the Cluster-Head (CH) or to change the entire cluster hierarchies periodically [69, 64, 62], others are interested in communication between nodes and among clusters [65, 66] or in cluster joining [70]. Hence, in this chapter, we adopt a cluster based architecture and we consider that the network is clustered and the CHs are defined using an appropriate clustering scheme.

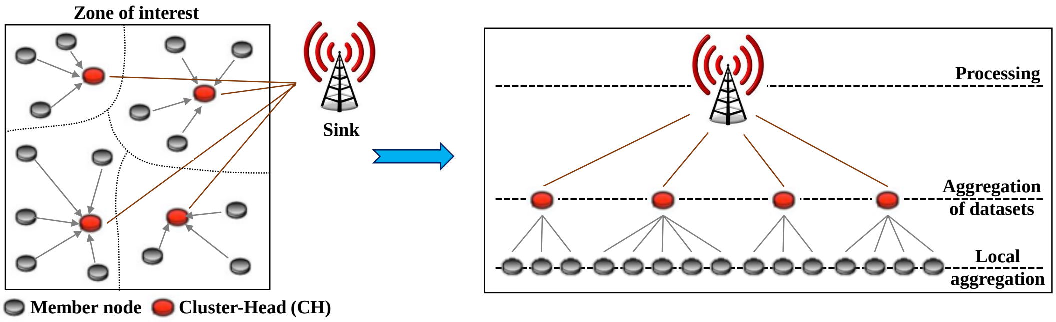


Figure 2.1: Data aggregation based on clustering network topology.

In Fig. 2.1, a cluster-based sensor network topology is presented. Each sensor node collects data in a periodic manner and sends it to the appropriate CH for aggregation and processing. We consider that sensor nodes operate at a fixed sampling rate where each one takes τ measures at each period and send the data set to the CH after local aggregation.

Remark 4: Sampling Rate

We also studied our aggregation techniques for nodes operating in different sampling rates. For more details, please refer to [71].

2.4/ Aggregation at sensor level: Local Aggregation

In WSN, measures collected by sensor nodes are highly dependent on the monitored feature (e.g. temperature, humidity, etc.). Hence, the monitored condition can slow down or speed up, resulting similar and redundant collected measurements. If each node sends the raw data to its CH and then to the sink, energy will be wasted and thus the network energy will be quickly depleted. In this section we present a non-complex local data aggregation technique to eliminate redundant data collected by the same node in the same period.

2.4.1/ Definitions and Notations

We consider that, each period p is divided into τ equal slots as follows: $p=[s_1, s_2, \ldots, s_\tau]$. At each slot s_j, each sensor S_i captures a new measurement m_{i_j}, then, it forms a vector of measures during the period p as follows: $M_i=[m_{i_1}, m_{i_2}, \ldots, m_{i_\tau}]$. Fig. 2.2 shows an example of periodic data collection where the sensor node S_i takes five measures (e.g. τ=5) at each period p_q (where $q \in [1,3]$) and sends its vector of collected data M_i=$\left[m_{i_1}, m_{i_2}, m_{i_3}, m_{i_4}, m_{i_5}\right]$ to the CH at the end of each period.

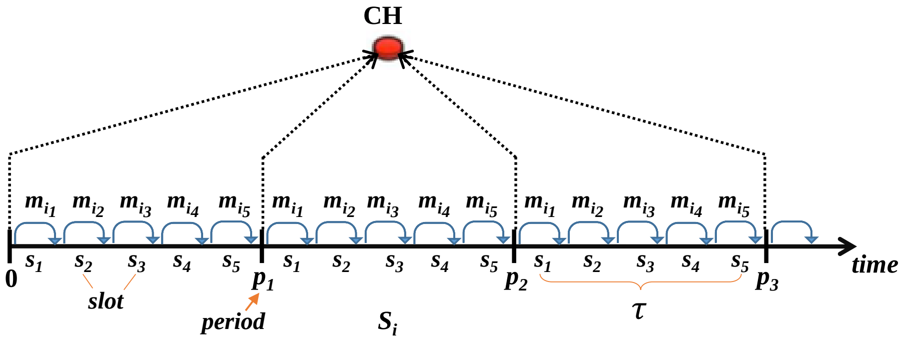

CH
m_{i_1} m_{i_2} m_{i_3} m_{i_4} m_{i_5}
m_{i_1} m_{i_2} m_{i_3} m_{i_4} m_{i_5}
m_{i_1} m_{i_2} m_{i_3} m_{i_4} m_{i_5}
s_1 s_2 s_3 s_4 s_5
s_1 s_2 s_3 s_4 s_5
s_1 s_2 s_3 s_4 s_5
0
p_1
p_2
p_3
time
slot
period
S_i
τ

Figure 2.2: Data collection in WSN.

As mentioned above, a data vector M_i formed by the sensor S_i may contain similar measurements, especially when the monitored condition varies slowly or when the slots are too short. In order to eliminate redundant values from the vector M_i, the sensor node S_i searches for measures similarities in the vector. Therefore, we define the $Similar$ function, i.e. $Similar(m_{i_j}, m_{i_k})$, to identify if the two measures m_{i_j} and m_{i_k} captured by the sensor S_i in the period p are similar or not as follows:

Definition 4: $Similar$ function

We define the $Similar$ function between two measurements captured by the same sensor node S_i at a period p as follows:

$$Similar(m_{i_j}, m_{i_k}) = \begin{cases} 1 & \text{if } \quad \left\| m_{i_j} - m_{i_k} \right\| \leqslant \varepsilon, \\ 0 & \text{otherwise.} \end{cases}$$

where m_{i_j} and $m_{i_k} \in M_i$ and ε is a threshold to be determined by the application. Furthermore, two measures are similar if and only if their $Similar$ function is equal to 1.

In order to save the integrity of the information, we define the weight of a measure as follows:

Definition 5: Measure's weight, $wgt(m_{i_j})$

The weight of a measurement m_{i_j} is defined as the number of similar measures (according to the $Similar$ function) to m_{i_j} in the same vector M_i.

Based on the notations defined above, we describe the local aggregation phase which is run by the nodes themselves at each period in the following manner: for each new captured measurement, a sensor node S_i 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, else it adds the new measure to the set and initializes its weight to 1[1]. After applying the local aggregation algorithm, S_i will transform the initial vector of measures, M_i, to a set of measures, M'_i, associated to their corresponding weights as follows: M'_i=$\{(m'_{i_1}, wgt(m'_{i_1})), (m'_{i_2}, wgt(m'_{i_2})), \ldots, (m'_{i_k}, wgt(m'_{i_k}))\}$, where $k \leqslant \tau$.

[1] Several methods could be applied like dichotomic search, compression, etc.

Illustrative example: let consider a vector of measures generated by the sensor S_i at the period p as follows: $M_i = [10, 10.2, 10.3, 11, 11.1, 12, 11.3, 10.4, 12.1, 12.4]$. By taking $\varepsilon = 0.5$, $Similar$ function will transform the vector M_i to the following set of measures: $M'_i = \{(10; 4), (11; 3), (12; 3)\}$, where 4, 3 and 3 are the weights of the measures 10, 11 and 12 respectively.

Based on the set M'_i, we provide the following definitions:

Definition 6: Cardinality of the set M'_i, $|M'_i|$

The cardinality of the set M'_i is equal to the number of elements in M'_i, i.e. $|M'_i| = k$.

Definition 7: Weighted Cardinality of the set M'_i, $wgt_c(M'_i)$

The weighted cardinality of the set M'_i is equal to the sum of all measures' weights in M'_i as follows: $wgt_c(M'_i) = \sum_{j=1}^{|M'_i|} wgt(m_{i_j})$.

At the end of each period p, each sensor node S_i will have a set M'_i with no redundant measures. The second step is to send it to the appropriate CH which, in turn, aggregates the data sets coming from different member nodes. In the next sections, we present three different aggregation methods at the CH level.

2.5/ Data Aggregation using sets similarity functions

At the end of each period, each CH will receive several sets of measurements and their weights from different nodes. The objective of the second aggregation level is to eliminate redundant data sets by identifying all pairs of sets whose similarities are above a given threshold. For this reason, we used in our work [11, 12] the Jaccard similarity function, which is one of the most widely accepted functions as it can support many other similarity functions and difficult to be satisfied [72]. The Jaccard similarity function returns a value in $[0, 1]$ where a higher value indicates that the sets share more similarities. Thus we can treat pairs of sets with high Jaccard similarity value as near duplicate to reduce the size of final data sets transmitted from the CH to the sink. A Jaccard similarity function between two sets M'_i and M'_j, generated respectively by the sensors S_i and S_j, can be formulated as follows [11]:

$$J(M'_i, M'_j) \geqslant t_J \Leftrightarrow |M'_i \cap_s M'_j| \geqslant \alpha = \frac{2 \times t_J \times \tau}{1 + t_J} \tag{2.1}$$

where t_J is the Jaccard threshold defined by the application itself and "$\cap_s$" is defined as:

Definition 8: $\cap_s$

Consider two sets of measurements M'_i and M'_j, then we define the overlap, $\cap_s$, between them as: $M'_i \cap_s M'_j = \{(m'_i, m'_j) \in M'_i \times M'_j$ with weight $wgt_{min}(m'_i, m'_j)\ and\ Similar(m'_i, m'_j) = 1\}$;

where $wgt_{min}(m'_i, m'_j) = min(wgt(m'_i), wgt(m'_j))$, the minimum value between the weights of m'_i and m'_j.

Then, in order to prevent the CH from enumerating and comparing every pair of sets which has a $O(n^2)$ number of comparisons, we proposed to use a prefix frequency filtering (PFF) technique [12]. The PFF technique works in the following two steps to find the pairs of similar sets:

- **Candidate pairs' generation:** in this step, the CH searches the candidates (which may be similar) sets for every data set. This step is based on the intuition that if all sets of measures are sorted by a global ordering, some fragments of them must share several common measures with each others in order to meet the Jaccard threshold similarity (t_J). Therefore, it first defines a prefix of length $|M'_i| - \lceil t_J \times |M'_i| \rceil + 1$ for every set M'_i. Then two sets M'_i and M'_j are candidates if and only if they share at least β measurements in their prefixes as shown in the following lemma [11]:

Lemma 1: Prefix Frequency Filtering

Assume that all the measures in the sets M'_i and M'_j are ordered in decreasing order of the measures weights. Let the p-prefix be the first p elements of M'_i. If $M'_i \cap_s M'_j \geqslant (2 \times t_J \times \tau)/(1 + t_J)$, then $p - M'_i \cap_s p - M'_j \geqslant \beta = \sum_{k=1}^{|p\text{-}M'_i|} wgt(m'_k) - \big((1 - t_J)/(1 + t_J)\big) \times \tau$ where $m'_k \in p\text{-}M'_i$.

Proof. We denote by $p\text{-}M'_i$ the prefix of the set M'_i and $r\text{-}M'_i$ the set of reminder measures where $M'_i = \{p\text{-}M'_i + r\text{-}M'_i\}$. We have:

$$\begin{aligned} M'_i \cap_s M'_j &= p\text{-}M'_i \cap_s M'_j + r\text{-}M'_i \cap_s M'_j \\ &= p\text{-}M'_i \cap_s p\text{-}M'_j + p\text{-}M'_i \cap_s r\text{-}M'_j + r\text{-}M'_i \cap_s M'_j \\ &\cong p\text{-}M'_i \cap_s p\text{-}M'_j + r\text{-}M'_i \cap_s M'_j \\ &\leqslant p\text{-}M'_i \cap_s p\text{-}M'_j + \sum_{k=1}^{|r\text{-}M'_i|} (wgt(m'_k \in r\text{-}M'_i)) \end{aligned}$$

In the second line we can omit the term $p\text{-}M'_i \cap_s r\text{-}M'_j$ because we have assumed that it is negligible compared to the other terms in the equation. Indeed, if the two sets are similar then the measures having highest weights must be in the prefix set and not in the reminder, which means that the overlapping between the $p\text{-}M'_i$ and $r\text{-}M'_j$ is almost empty. From the above equations and equation (2.1)(similarity condition) we can deduce:

$$\frac{2 \times t_J \times \tau}{1 + t_J} \leqslant p\text{-}M'_i \cap_s p\text{-}M'_j + \sum_{k=1}^{|r\text{-}M'_i|} wgt(m'_k \in r\text{-}M'_i) \tag{2.2}$$

From the following equation:

$$\sum_{k=1}^{|p\text{-}M'_i|} wgt(m'_k \in p\text{-}M'_i) + \sum_{k=1}^{|r\text{-}M'_i|} wgt(m'_k \in r\text{-}M'_i) = \tau \tag{2.3}$$

We obtain:

$$p\text{-}M'_i \cap_s p\text{-}M'_j \geqslant \sum_{k=1}^{|p\text{-}M'_i|} wgt(m'_k \in p\text{-}M'_i) - \frac{1 - t_J}{1 + t_J} \times \tau \tag{2.4}$$

The lemma is proved. □

Based on the lemma 1, the CH calculates the overlap between the prefix of each pair of sets; the two sets are considered a candidate pair if their calculated overlap is greater than β.

- **Candidates' verification:** once all the candidates pairs are found, the CH verifies the Jaccard similarity for each one in the second step. The two sets in a candidate pair are considered similar if their similarity is greater than the Jaccard threshold t_J.

Algorithm 4 describes the PFF technique to find similar sets. Briefly, the CH searches similar measures between prefixes of every pair of sets using the $Similar$ function (lines 3-21). Then, it assumes that the two sets are a candidate pair only if the overlap between their prefixes is greater than the score determined at lemma 1 (line 22). Finally, the two sets are considered similar if the overlap between their measures is greater than the Jaccard threshold (lines 23–25). For more details about this algorithm, please refer to Algorithm 4 in [11].

Algorithm 4: PFF Algorithm.

```
Require: Set of measures' sets $M' = \{M'_1, M'_2 ... M'_n\}$, $t_J$.
Ensure: All pairs of sets $(M'_i, M'_j)$, such that $J(M'_i, M'_j) \geqslant t_J$.
 1: $S \leftarrow \varnothing$
 2: $I_i \leftarrow \varnothing (1 \leqslant i \leqslant$ total number of measures in the prefixes of all sets)
 3: for each set $M'_i \in M'$ do
 4:     $p \leftarrow |M'_i| - \lceil t_J \times |M'_i| \rceil + 1$
 5:     $F_s \leftarrow$ empty map from set id to int
 6:     $sumFreq \leftarrow 0$
 7:     for $k \leftarrow 1$ to $p$ do
 8:         $sumFreq \leftarrow sumFreq + wgt(m'_k)$, where $m'_k \in p\text{-}M'_i$
 9:     end for
10:     for $k \leftarrow 1$ to $p$ do
11:         $w \leftarrow M'_i[k]$
12:         if ($I_{w_s}$ exists such that $Similar(w, w_s) = 1$) then
13:             for each Measurement $(M'_j[l]), wgt(M'_j[l]) \in I_{w_s}$ do
14:                 $F_s[M'_j] \leftarrow F_s[M'_j] + wgt_{min}(M'_i[k], M'_j[l])$
15:             end for
16:             $I_{w_s} \leftarrow I_{w_s} \cup \{p - M'_i\}$
17:         else
18:             create $I_w$
19:             $I_w \leftarrow I_w \cup \{p - M'_i\}$
20:         end if
21:     end for
22:     for each $M'_j$ such that $Fs[M'_j] \geqslant sumFreq - ((1 - t_J)/(1 + t_J)) \times \tau$ do
23:         if $J(M'_i, M'_j) \geqslant \alpha$ then
24:             $S \leftarrow S \cup \{(M'_i, M'_j)\}$
25:         end if
26:     end for
27: end for
28: return $S$
```

Remark 5: Some optimizations

Although filtering approaches reduce the number of comparisons between the received sets of measures, the number of candidate sets surviving after this phase still important. In order to decrease the data latency, we provided several optimizations of the PFF technique based on the suffix filtering [73, 74]. Furthermore, the computation of the jaccard similarity between two candidates sets can be very complex, especially when it comes to sensor networks where measures' sets can have ten hundreds or thousands elements. Therefore, to continue filtering out further candidate sets we proposed in [12] a new frequency based constraint in the verification phase. In doing so, we can also reduce the overhead of the jaccard similarity computation.

2.6/ Data Aggregation using K-means and ANOVA Model

Studying the variance between measurements in the data sets is another way of finding nodes that generate redundant data sets. In this section, our objective is to present briefly our technique proposed in [13], based on the k-means algorithm and the one way Anova model with Bartlett test. Indeed, the one-way Anova model is used to identify if the variance (R) between measures in a group of data sets is significant or not. R can be calculated in different manners depending on the applied statistical tests. In [75], we used the Anova model and compared different statistical tests (Fisher, Tukey and Bartlett) in order to detect all pairs of nodes with identical behavior which generate redundant data logs or sets. As the Bartlett test gives the best results, in this chapter, it is used with the k-means Algorithm and compared to the two other methods. Once R is calculated, the sets are considered duplicated if R is less than a threshold T (fixed by the test) for some desired false-rejection probability (risk α).

R is calculated according to the Bartlett test as follows [13]:

$$R = \frac{(\tau - 1)(n \times \ln(\sigma_p^2) - \sum_{j=1}^{n} \ln(\sigma_j^2))}{\lambda} \tag{2.5}$$

where n is the number of total sets, τ the total number of collected measures during the period p, σ_j^2 is the variance of the set M'_j and:

$$\lambda = 1 + \frac{(n+1)}{3 \times n \times (\tau - 1)} \tag{2.6}$$

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

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

Thus the decision is based on the following rule:

- if $R > T$, the variance between the sets is significant thus the sets are not considered redundant.

- if $R \leqslant T$, the variance between the sets is not significant.

In order to apply the Anova model over the groups of sets, we used the k-means algorithm to classify the sets in groups based on the means of these sets. Then, we proposed a new initialization method to find, dynamically, the optimal number of groups (K) in k-means [13]. The proposed method divides a parent group into $\lfloor \sqrt{n/2} \rfloor$ children groups, where n is the number of total sets at each period, every time the Anova model is not satisfied. Finally, the CH sends one data set from each group with the IDs of all the sensors in this group to the sink.

Algorithm 5 describes the k-means algorithm adopted by the Anova model and the Bartlett test, which we call it the **KAB** technique. First, it starts, as explained previously, by grouping all the received sets at the initial same group (lines 4 to 6). Then, it searches the variance between measurements in all the sets in the initial group, using the Anova model and the Bartlett test (lines 9 and 10). If the test's result indicates a low variance between the sets then, the algorithm considers this group as a final group and it puts it in the list of final groups (lines 11, 12 and 13). Else, it divides the initial group in K sub groups by applying the k-means algorithm (line 15). Once the final groups are obtained, CH sends only one useful information to the sink, e.g. the data set with the highest cardinality, and the IDs for the sensors that generate redundant data sets (lines 19 to 22).

Algorithm 5: K-means Adopted to Variance Study

```
Require: Set of measures' sets $M' = \{M'_1, M'_2 ... M'_n\}$, $K$.
Ensure: List of selected sets, $L$.
    $C \leftarrow \emptyset$ // list of all final groups
 2: $Q \leftarrow \emptyset$ // a temporary list of groups
    $C_1 \leftarrow \emptyset$
 4: for each set $M'_i \in M'$ do
      $C_1 \leftarrow C_1 \cup \{M'_i\}$
 6: end for
    $Q \leftarrow Q \cup \{C_1\}$
 8: repeat
      compute $R$ for $C_i$ based on Equation 2.5
10:   find $T$
      if $R \leqslant T$ then
12:     $C \leftarrow C \cup \{C_i\}$
        remove $C_i$ from $Q$
14:   else
        $Q \leftarrow Q \cup$ k-means($C_i, K$)
16:   end if
    until no cluster $C_i \in Q$
18: $L \leftarrow \emptyset$
    for each cluster $C_i \in C$ do
20:   consider $|M'_j| > |M'_{j*}|$; where $M'_{j*} \in C_i - \{M'_j\}$
      $L \leftarrow L \cup \{(M'_j, ID(M'_j) \cup ID(M'_{j*}))\}$
22: end for
    return $L$
```

2.7/ Data Aggregation using distance functions

In this section, we propose another technique to search redundant data sets generated by the sensors using the distance functions. Distance functions are an important method that can find duplicated data sets by searching dissimilarities between these sets. Hence, a great number of distance functions have been proposed in the literature [76]. In this work, we are interested in two distance functions that are widely used in various domains: Euclidean and Cosine distances.

Let us consider two data sets M'_i and M'_j, generated by the sensor nodes S_i and S_j respectively, at the period p as follows: $M'_i = \{(m'_{i_1}, wgt(m'_{i_1})), (m'_{i_2}, wgt(m'_{i_2})), \ldots, (m'_{ik_i}, wgt(m'_{ik_i}))\}$ and $M'_j = \{(m'_{j_1}, wgt(m'_{j_1})), (m'_{j_2}, wgt(m'_{j_2})), \ldots, (m'_{jk_j}, wgt(m'_{jk_j}))\}$ where $|M'_i| = k_i$ and $|M'_j| = k_j$. Therefore, M'_i and M'_j are considered redundant if the calculated distance between them is less than a threshold (t_d) as follows:

$$Dist(M'_i, M'_j) \leqslant t_d$$

However, two issues must be considered when using distance functions in the context where the measures are assigned with their weights: **1)** Calculating the distance between two data sets with different cardinalities, e.g. k_i and k_j, and **2)** integrating the weights when calculating the distance between sets. To face these challenges, we propose to use the threshold ε, introduced in the $Similar$ function (cf. Section 2.4), when computing the distance between the sets.

In order to find the distance between two sets M'_i and M'_j, the first step consists in dividing each set into two parts: overlap and remained. The overlap part of the set M'_i (resp. M'_j) contains measures that are similar to those in M'_j (resp. M'_i) while the remained part contains the remaining measures of M'_i (resp. M'_j). Subsequently, the overlap part between two sets has already been defined in Definition 8, i.e. $M'_i \cap_s M'_j$, while the remained part in each set is defined as follows:

Definition 9: Remained part of M'_i, M'_{i_r}

Consider two sets of sensor measures M'_i and M'_j. We define the remained part M'_{i_r} (respectively M'_{j_r}) as all the measures in M'_i (respectively M'_j) minus the measures in the overlap part of M'_i (respectively M'_j) as follows:

$$\begin{cases} & M'_{i_r} = M'_i \ominus (M'_i \cap_s M'_j) \\ and & \\ & M'_{j_r} = M'_j \ominus (M'_i \cap_s M'_j) \end{cases}$$

Where $\ominus$ is a new operation defined as:

Definition 10: Minus Operation, $\ominus$

We define the minus operation, $M'_i \ominus M'_j$, between two sets M'_i and M'_j as all the measures in M'_i and not in M'_j as follows:
$M'_i \ominus M'_j = \{m'_i \in M'_i$, with $wgt(m'_i) = wgt(m'_i) - wgt(m'_j)$ for $m'_j \in M'_i \cap_s M'_j$ and $Similar(m'_i, m'_j) = 1\}$

In order to compute the distance between M'_i and M'_j, we must transform M'_{i_r} (respectively M'_{j_r}) into a vector as follows:

$$vM'_{i_r} = [\underbrace{m'_{i_1},\ldots,m'_{i_1}}_{wgt(m'_{i_1})\text{ times}},\underbrace{m'_{i_2},\ldots,m'_{i_2}}_{wgt(m'_{i_2})\text{ times}},\ldots,\underbrace{m'_{i_{k_i}},\ldots,m'_{i_{k_i}}}_{wgt(m'_{i_{k_i}})\text{ times}}]$$

2.7.1/ Euclidean distance

The Euclidean distance is used in many applications and domains, such as computer vision and prevention of identity theft [77].

In general, the Euclidean distance (E_d) between two data sets M_i and M_j, before applying the local aggregation, is given by:

$$E_d(M_i, M_j) = \sqrt{\sum_{k=1}^{\tau}(m_{i_k} - m_{j_k})^2} \quad \text{where } m_{i_k} \in M_i \text{ and } m_{j_k} \in M_j \tag{2.7}$$

Thus, M_i and M_j are said to be redundant if $E_d(M_i, M_j) \leqslant t_d$, where t_d is a threshold determined by the application.

After applying the local aggregation phase, we consider that M_i and M_j are respectively transformed into M'_i and M'_j. Therefore, we calculate the Euclidean distance between M'_i and M'_j as follows:

Lemma 2: Euclidean distance computation

The Euclidean distance between two sets of data is equal only to the distance between measures in the remained parts of M'_i and M'_j, i.e. vM'_{i_r} and vM'_{j_r} respectively.

$$E_d(M'_i, M'_j) = E_d(vM'_{i_r}, vM'_{j_r}) = \sqrt{\sum_{k=1}^{|vM'_{i_r}|}(m'_{i_k} - m'_{j_k})^2} \tag{2.8}$$

where $m'_{i_k} \in vM'_{i_r}$ and $m'_{j_k} \in vM'_{j_r}$

Proof.

$$\begin{aligned}
E_d(M'_i, M'_j) &= \sqrt{(M'_i - M'_j)^2} \\
&= \sqrt{\Big(\big(M'_i \cap_s M'_j + vM'_{i_r}\big) - \big(M'_i \cap_s M'_j + vM'_{j_r}\big)\Big)^2} \\
&= \sqrt{\Big(\big(M'_i \cap_s M'_j - M'_i \cap_s M'_j\big) + \big(vM'_{i_r} - vM'_{j_r}\big)\Big)^2} \\
&= \sqrt{(vM'_{i_r} - vM'_{j_r})^2} \\
&= \sqrt{\textstyle\sum_{k=1}^{|vM'_{i_r}|}(m'_{i_k} - m'_{j_k})^2} \quad \text{where } m'_{i_k} \in vM'_{i_r} \text{ and } m'_{j_k} \in vM'_{j_r}
\end{aligned}$$

□

In the above proof, we consider that the Euclidean distance between the measures in the overlap is equal to zero because they are redundant.

2.7.2/ Cosine distance

Cosine distance is a measure of dissimilarity between two vectors that measures the cosine of the angle between them. This kind of dissimilarity has been used widely in many aspects, such as the anomaly detection in web documents [78] and medical diagnosis [79]. Depending on the angle between the vectors, the resulting dissimilarity ranges from -1 meaning exactly the opposite, to 1 meaning exactly the same.

The Cosine distance (C_d) between two sets M_i and M_j, before applying local aggregation, is given by:

$$C_d(M_i, M_j) = 1 - \frac{\sum_{k=1}^{\tau}(m_{i_k} \times m_{j_k})}{\sqrt{\sum_{k=1}^{\tau} m_{i_k}^2} \times \sqrt{\sum_{k=1}^{\tau} m_{j_k}^2}} \quad \text{where } m_{i_k} \in M_i \text{ and } m_{j_k} \in M_j. \tag{2.9}$$

Thus, M_i and M_j are redundant if $C_d(M_i, M_j) \leqslant t_d$.

Lemma 3: Cosine distance computation

The Cosine distance between two sets of data M'_i and M'_j can be computed as follows:

$$C_d(M'_i, M'_j) = 1 - \frac{A + \sum_{k=1}^{|vM'_{i_r}|}(m'_{i_{rk}} \times m'_{j_{rk}})}{\sqrt{A + \sum_{k=1}^{|vM'_{i_r}|} m'^2_{i_{rk}}} \times \sqrt{A + \sum_{k=1}^{|vM'_{j_r}|} m'^2_{j_{rk}}}} \tag{2.10}$$

where $A = \sum_{k=1}^{|M'_i \cap_s M'_j|} \left(wgt_{min}(m'_{i_k}, m'_{j_k}) \times m'^2_{i_k}\right)$

Proof.

$$\begin{aligned}
C_d(M'_i, M'_j) &= 1 - \frac{M'_i \times M'_j}{\sqrt{M'^2_i} \times \sqrt{M'^2_j}} \\
&= 1 - \frac{\left(M'_i \cap_s M'_j + vM'_{i_r}\right) \times \left(M'_i \cap_s M'_j + vM'_{j_r}\right)}{\sqrt{(M'_i \cap_s M'_j)^2 + vM'^2_{i_r}} \times \sqrt{(M'_i \cap_s M'_j)^2 + vM'^2_{j_r}}} \\
&= 1 - \frac{(M'_i \cap_s M'_j)^2 + (vM'_{i_r} \times vM'_{j_r})}{\sqrt{(M'_i \cap_s M'_j)^2 + vM'^2_{i_r}} \times \sqrt{(M'_i \cap_s M'_j)^2 + vM'^2_{j_r}}} \\
&= 1 - \frac{A + \sum_{k=1}^{|vM'_{i_r}|}(m'_{i_{rk}} \times m'_{j_{rk}})}{\sqrt{A + \sum_{k=1}^{|vM'_{i_r}|} m'^2_{i_{rk}}} \times \sqrt{A + \sum_{k=1}^{|vM'_{j_r}|} m'^2_{j_{rk}}}}
\end{aligned}$$

where $A = \sum_{k=1}^{|M'_i \cap_s M'_j|} \left(wgt_{min}(m'_{i_k}, m'_{j_k}) \times m'^2_{i_k}\right)$

□

2.7.3/ Distance normalization

In general, each distance function has its own method to calculate the distance between data sets. For instance, straight-line distance in Euclidean distance and the angle between data sets in Cosine distance. Therefore, normalization becomes essential to scale the distance between data sets into the range $[0, 1]$ to have thus the same variation between sets before comparing them. Hence, Gaussian normalization has been considered as a better approach to normalize data sets [80]. Consequently, in our approach, data sets sent by the sensor nodes to the CH are normalized using Gaussian normalization. Once the CH receives the data sets at each period, it calculates first the distance, Euclidean or Cosine, for each pair of sets as follows:

$$d = \{d_1(M'_1, M'_2), d_2(M'_1, M'_3), \dots, d_{\frac{n\times(n-1)}{2}}(M'_{n-1}, M'_n)\}$$

where n is the number of total sets and $\frac{n\times(n-1)}{2}$ is the number of all possible distances. Then, it normalizes the returned distance values using the following Gaussian normalization equation:

$$d'_i = \frac{d_i - \overline{Y}}{6 \times \sigma} + \frac{1}{2} \tag{2.11}$$

where $\overline{Y}$ is the mean of all distances and σ is the standard deviation of pairwise distance over all data. $\overline{Y}$ and σ are calculated as follows:

$$\overline{Y} = \frac{\sum_{k=1}^{|d|} d_k}{|d|} \quad \text{and} \quad \sigma = \sqrt{\frac{\sum_{k=1}^{|d|}(d_i - \overline{Y})^2}{|d|}}$$

where $|d| = \frac{n\times(n-1)}{2}$.

After normalizing all pairwise distances, the CH will form the distance normalization vector between each pair of sets as follows:

$$d' = \{d'_1(M'_1, M'_2), d'_2(M'_1, M'_3), \dots, d'_{\frac{n\times(n-1)}{2}}(M'_{n-1}, M'_n)\}.$$

2.7.4/ Distance-based aggregation algorithm

Algorithm 6 describes how the CH finds redundant sets of measures generated by sensors then how it selects, among them, data sets to be sent to the sink. After normalizing data sets based on Equation 2.11 (lines 2 to 11), the CH considers that two sets are redundant if the normalized distance between them is less than the threshold t_d (line 12 and 13). $Dist$ function in line 5 will be replaced by the distance function (Euclidean, Cosine, Camberra, etc. [14]) and can be calculated based on Equations 2.8 and 2.10 for Euclidean and Cosine respectively. Then, for each pair of redundant set, the CH chooses the one having the highest cardinality (line 18), then it adds it to the list of sets to be sent to the sink (line 19). After that, it removes all pairs of redundant sets that contain M'_i or M'_j from the set of pairs (which means it will not check them again). Finally, the CH assigns to each set its weight (line 21) when sending it to the sink.

Algorithm 6: Distance-based Redundancy Searching Algorithm

```
Require: Set of measures' sets $M' = \{M'_1, M'_2 ... M'_n\}$, $t_d$.
Ensure: List of sent sets, $L$.
    $S \leftarrow \emptyset$
    $d \leftarrow \emptyset$ // list of pairwise distance
    for each set $M'_i \in M'$ do
        for each set $M'_j \in M'$ such that $j > i$ do
            compute $Dist(M'_i, M'_j)$
            $d \leftarrow d \cup \{Dist(M'_i, M'_j)\}$
        end for
    end for
    compute $\overline{Y}$ and $\sigma$ for $d$
    for each $d_i \in d$ do
        $d'_i = ((d_i - \overline{Y})/(6 \times \sigma)) + 0.5$
        if $d'_i \leqslant t_d$ then
            $S \leftarrow S \cup \{(M'_i, M'_j)\}$
        end if
    end for
    $L \leftarrow \emptyset$
    for each pair of sets $(M'_i, M'_j) \in S$ do
        Consider $|M'_i| \geqslant |M'_j|$
        $L \leftarrow L \cup \{M'_i\}$
        Remove all pairs of sets containing one of the two sets $M'_i$ and $M'_j$
        $wgt(M'_i)$ = number of removed pairs + 1
    end for
    return $L$
```

2.8/ Performance Evaluation

In order to study the efficiency of our proposed techniques, we developed a Java based simulator that executes with real data collected from 46 sensors deployed in the Intel Berkeley Research Lab [81]. Mica2Dot sensors with weather boards collect timestamped topology information, along with humidity, temperature, light and voltage values once every 31 seconds. The objective of these experiments is to confirm that the proposed data agregation methods can successfully achieve desirable results for energy conservation, data latency and data accuracy in different monitoring applications. Data were collected using TinyDB in-network query processing system built on the TinyOS platform. In our experiments, we used a file that includes a log of about 2.3 million readings collected from these sensors. For the sake of simplicity, in this chapter we show only the results for the temperature[2]. We assume that all nodes send their data to a common CH placed at the center of the Lab. First, each node periodically reads real measures while applying the local aggregation. At the end of this step, each node sends its set of measures with weights to the CH which in its turn aggregates them using the three proposed methods.

We evaluated the performance of the techniques using the following parameters

(a) the threshold ε, defined in $Similar$ function. Its value is varying as follows: 0.03, 0.05, 0.07, 0.1.

(b) τ, the total number of measures collected by each sensor node during a period. It

[2] the others were done by the same manner and gave similar results

will be varied as follows: 200, 500 and 1000.

(c) the distance threshold t_d while assigning the following values: 0.35, 0.4, 0.45 and 0.5.

(d) the threshold t_J of the Jaccard similarity function is fixed to 0.75.

(e) α, the false-rejection probability in the Anova model, is fixed to 0.01.

2.8.1/ Data aggregation at the sensor level

During the local aggregation, each sensor node searches the similarity between measures captured at each period and assigns for each measure its weight. Therefore, the result of the aggregation in this phase depends on the chosen threshold ε, the number of the collected measures in a period τ and the changes in the monitored condition. Fig. 2.3 shows the percentage of remaining data, or aggregated data, which will be sent to the CH, with and without applying the local aggregation phase at the sensor level. At each period, the amount of data collected by each sensor is reduced at least by 77% (and up to 94%) after applying the aggregation phase. Otherwise, the sensor node sends all the collected data, e.g. 100%, without applying the aggregation phase. Therefore, our technique can successfully eliminate redundant measures at each period and reduce the amount of data sent to the CH. We can also observe that, with the local aggregation phase, data redundancy among data increases when τ or ε increases. This is because the $Similar$ function will find more similar measures to be eliminated in each period.

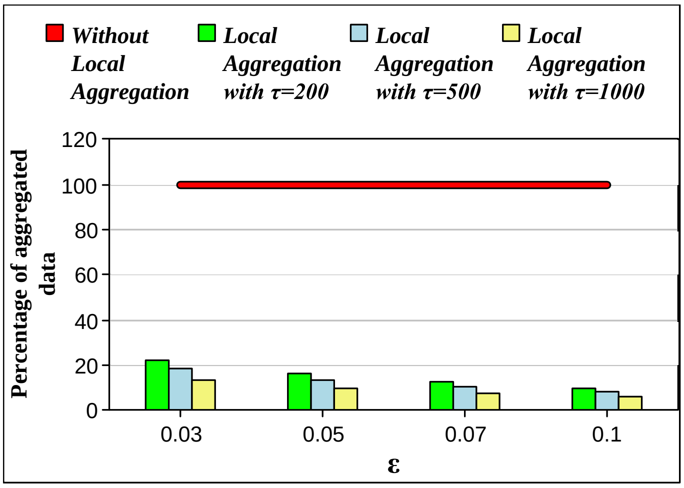
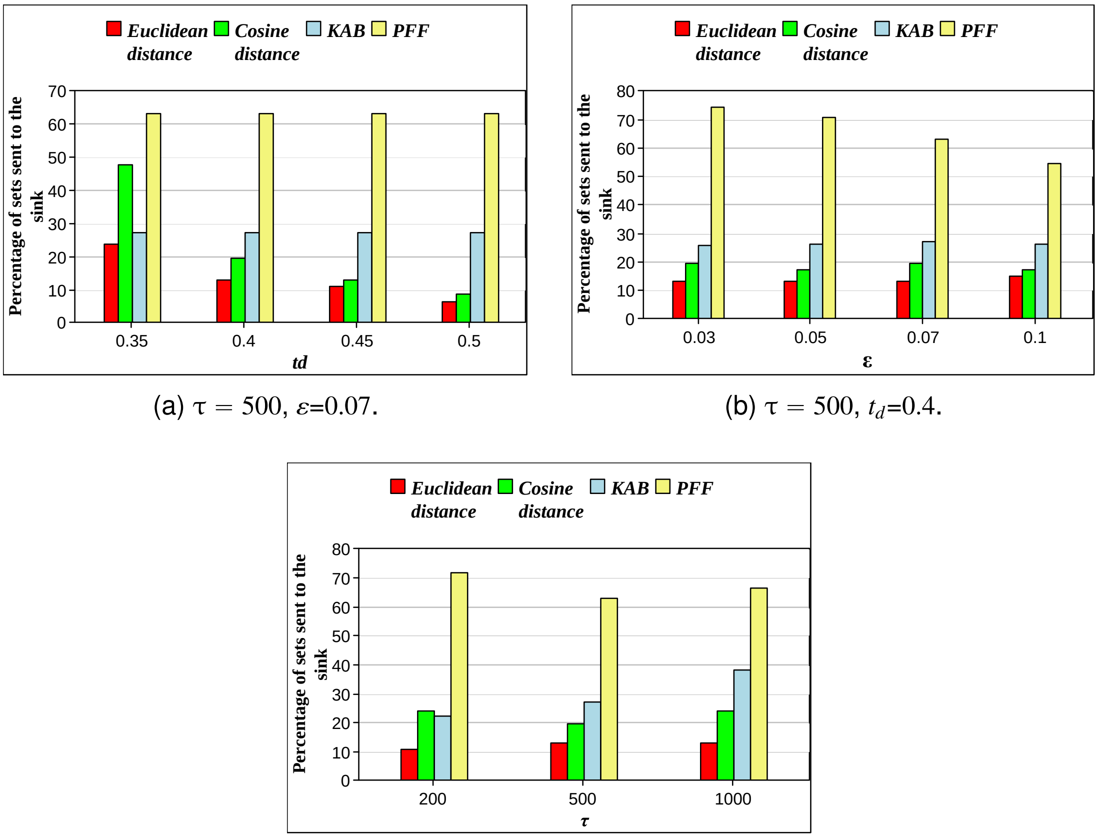


Figure 2.3: Percentage of remaining data after applying local aggregation.

2.8.2/ Data sets reduction

In this section, we show how the CH is able to eliminate redundant sets at each period before sending them to the sink.

Fig. 2.4 shows the the percentage of the remaining sets that will be sent to the sink after eliminating the redundancy when applying Euclidean and Cosine distances, K-means with Anova model (KAB) and Prefix-Frequency Filtering (PFF) techniques respectively. First, we fixed τ and ε and we varied t_d as shown in Fig. 2.4a, then we fixed τ and t_d and varied ε as shown in Fig. 2.4b and, finally, we fixed ε and t_d and varied τ as shown in Fig. 2.4c. Generally, the obtained results are dependent on the number of the redundant sets. We notice that Euclidean and Cosine distances allow the CH to eliminate more redundant sets, except when t_d is small (e.g. $t_d = 0.35$ in Fig. 2.4a), at each period compared to PFF and KAB techniques.



(a) $\tau = 500$, ε=0.07.

(b) $\tau = 500$, t_d=0.4.

(c) ε=0.07, t_d=0.4.

Figure 2.4: Percentage of sets sent to the sink after each period.

Several observations can be made based on the obtained results in Fig. 2.4:

- Using the Euclidean distance function allows the best data reduction comparing to the Cosine function, Kab and PFF. It means that more redundant data sets were found. This is due to the Euclidean distance condition which is more easier to be satisfied than other conditions.

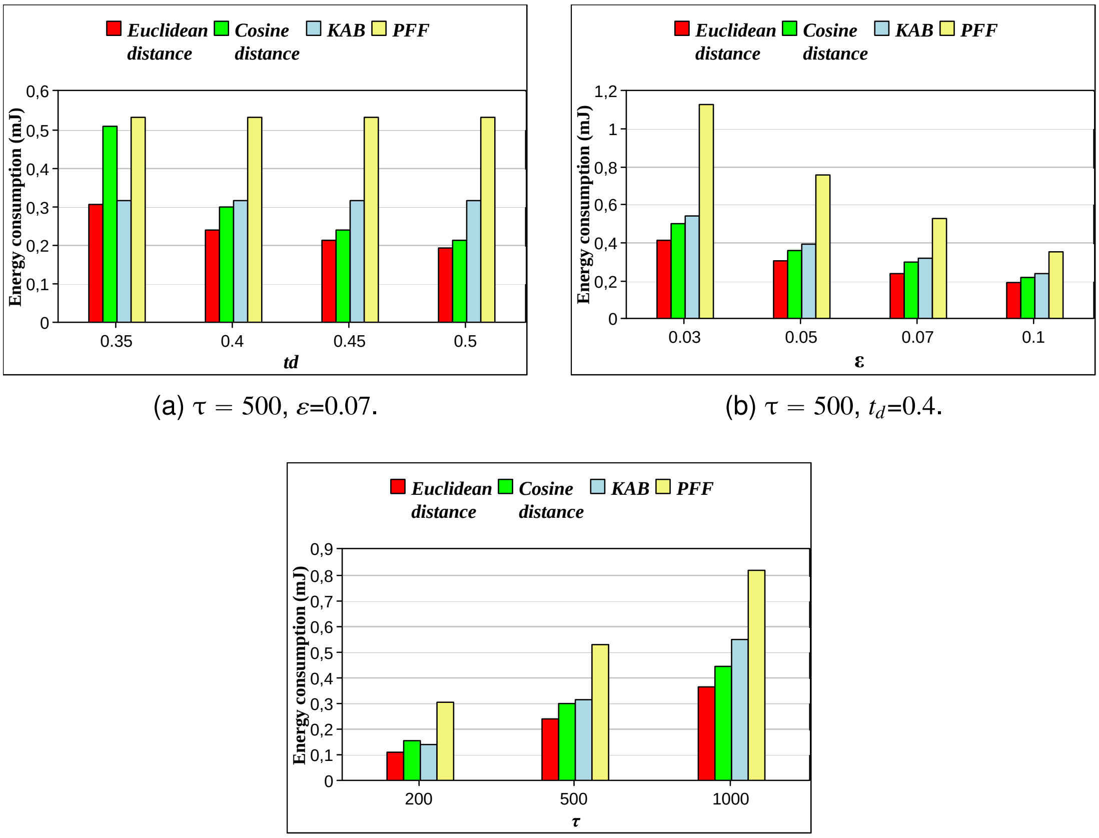
- The distance functions allow the CH to send 15% to 61% less of sets to the sink compared to PFF, due to the flexibility of distances regarding the redundancy compared to similarity functions. Similarly, KAB gives better results (36% to 49% less of sent sets) compared to PFF in all the cases. This is because KAB searches redundant sets in groups then sends one set of each group to the sink while PFF searches them by pairs.

- The percentage of sets sent to the sink in Euclidean and Cosine distances decreases when t_d increases (Fig. 2.4a) while it is almost fix when ε or τ increases (Fig. 2.4b and 2.4c).

- PFF sends less percentage of sets to the sink when ε increases while it is almost fix when using KAB technique (Fig. 2.4b). This is because the variance condition in Anova is not dependent on ε like the Jaccard function in PFF.

2.8.3/ Energy consumption at the CH level

Fig. 2.5 shows the energy consumption comparison at the CH when using distances method, KAB and PFF techniques, in function of t_d in Fig. 2.5a, of ε in Fig. 2.5b and of τ in Fig. 2.5c. Regarding the obtained results in Fig. 2.4, it is obvious that the distances method gives the best energy consumption results. Subsequently, Euclidean distance can reduce the energy consumed in CH up to 39% and up to 64% compared to the amount of energy consumed using KAB and PFF respectively. Otherwise, Cosine distance can reduce up to 33% and 60% of the energy consumption in the CH compared to that consumed using KAB and PFF respectively. On the other hand, energy consumption in the CH is reduced, using KAB technique, from 32% to 54% compared to that consumed using PFF.



(a) $\tau = 500$, ε=0.07.

(b) $\tau = 500$, t_d=0.4.

(c) ε=0.07, t_d=0.4.

Figure 2.5: Energy consumption at the CH level.

From Fig. 2.5 we can notice that :

- The Euclidean distance method decreases the energy consumption in the CH from 9% to 40% compared to the Cosine distance.

- Using Euclidean and Cosine distances, the CH conserves more energy when t_d increases (Fig. 2.5a).
- The PFF technique reduces the energy consumption in the CH when ε increases (Fig. 2.5b).

2.8.4/ Data latency and Execution time

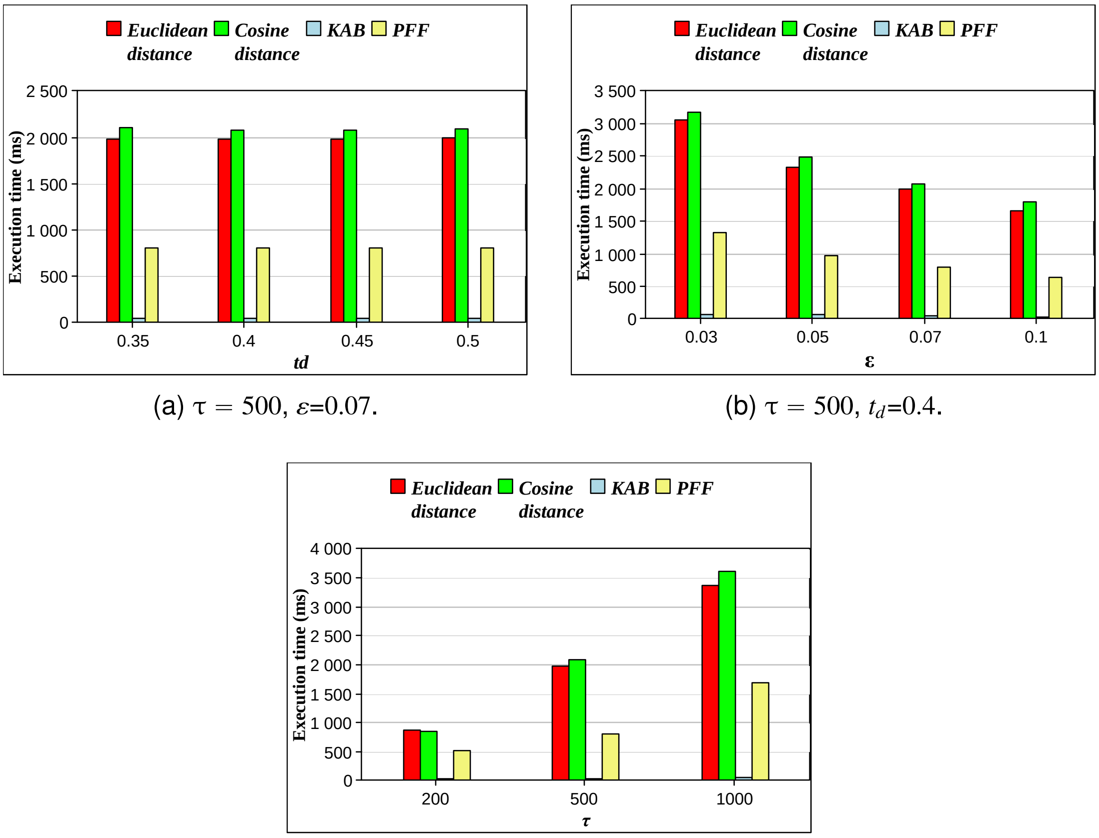
In this section, we compare the execution time required for the three data aggregation methods when varying t_d, ε and τ respectively (Fig. 2.6). The execution time is dependent on the normalization process of data sets in distances method, on the number of iterative loops in k-means algorithm used in KAB technique, and on the number of candidates generated in PFF. The obtained results show that KAB significantly outperforms the other methods in terms of computation, in all cases. This is because, searching the groups of redundant sets in KAB requires less computation time compared to the computation time required for comparison by pairs used in distances and similarity methods. Consequently, KAB can accelerate the execution time at the CH from 23 to 57 times compared to distances and from 14 to 26 compared to PFF. On the other hand, PFF can accelerate the execution time at the CH twice faster than distances method; the reason for that is the normalization used in Euclidean and Cosine distances which needs to calculate all distances between pairs of sets while PFF only searches the similarity between the generated candidate pairs.

Several observations can be made based on the results shown in Fig. 2.6:

- The Euclidean distance decreases the execution time at the CH more than the Cosine distance. This is due to the complexity of the calculation of Cosine distance (Equation 2.10) compared to the Euclidean distance (Equation 2.8).
- The execution time required for both distances is almost fix when varying t_d (Fig. 2.6a). This is because both distances must normalize all data sets independently from t_d value.
- The data latency at the CH is optimized when ε increases, in all methods (Fig. 2.6b). This is because the cardinality of the data sets decreases when ε increases thus the computation between sets decreases as well.
- The CH requires, with the three aggregation methods, more execution time when τ increases (Fig. 2.6c). This is because the cardinality of sets increases which requires more time to be processed.

2.8.5/ Data accuracy: Aggregation error

Data accuracy is an important factor in WSNs which represents the measures “loss rate”. In our simulation, data accuracy has been evaluated based on the percentage of loss measures at the CH. We compute the ratio between the number of the collected measures by all sensor nodes whose values (or similar values) do not reach the sink, over the whole collected measures. Fig. 2.7 shows the results of data accuracy for the three data aggregation techniques for different values of t_d, ε and τ. We can notice that PFF gives



(a) $\tau = 500$, ε=0.07.

(b) $\tau = 500$, t_d=0.4.

(c) ε=0.07, t_d=0.4.

Figure 2.6: Execution time at the CH.

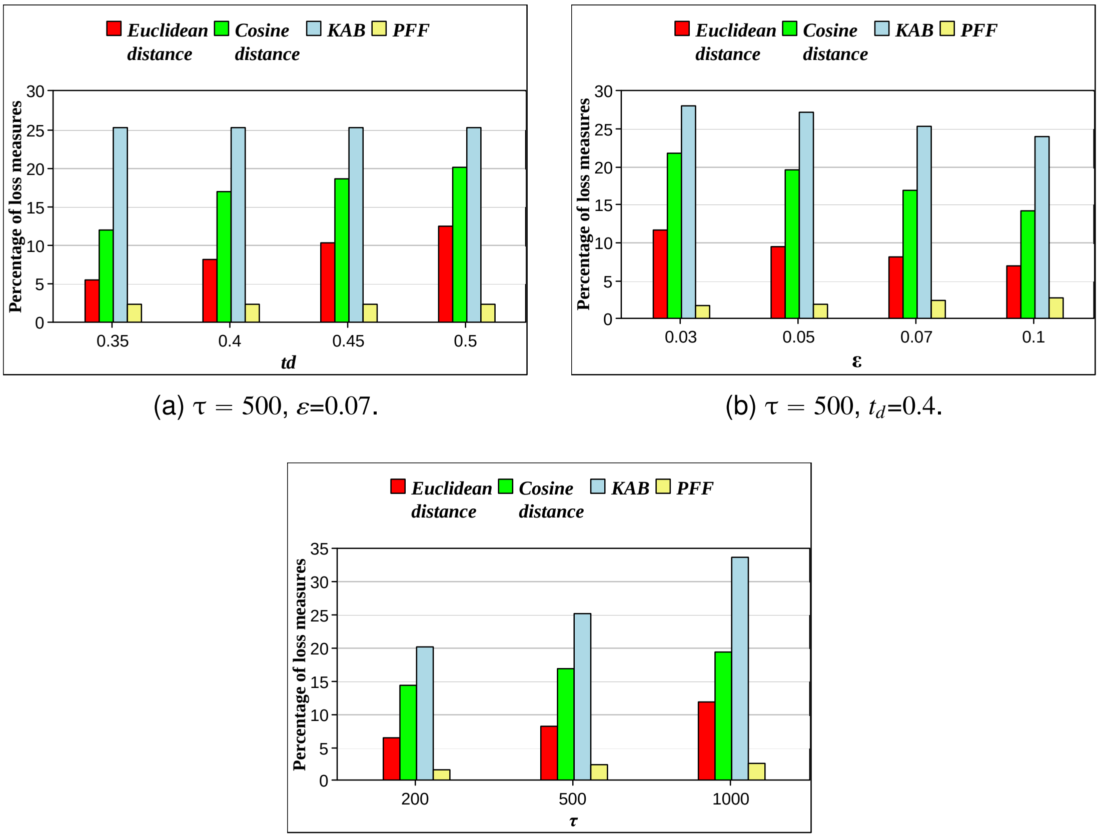
the best results for data accuracy, 2.81% in the worst case, compared to the Euclidean (up to 12.52%) and Cosine (up to 21.75%) distances and KAB technique (up to 33.8%). The reason for this is that the Jaccard function used in PFF is a strong constraint regarding the loss measures compared to distance and variance constraints which are more flexible. We can also notice that, Euclidean and Cosine distances conserve the integrity of the information more than the KAB technique. Indeed, KAB sends one set among a group of sets to the sink which increases the loss of measures. It is important to know that with KAB technique, the objective is to send the minimum amount of data to the sink, which allows decision makers to take the correct decision based on the received information.

The following observations can be made based on the results obtained in Fig. 2.7:

- The Euclidean distance conserves the integrity of data more than Cosine distance in all cases. This is due to the equation of Cosine distance which eliminates the sets that have high cardinality.

- The loss of measures using Euclidean and Cosine distances increases when t_d increases (Fig. 2.7a). This is because the CH eliminates more sets when t_d increases (see results in Fig. 2.4).

- The data accuracy, in distances and KAB techniques, increases when ε increases (Fig. 2.7b) or τ decreases (Fig. 2.7c). On the other hand, using PFF, the data accuracy decreases when ε or τ increases.



(a) $\tau = 500$, ε=0.07.

(b) $\tau = 500$, t_d=0.4.

(c) ε=0.07, t_d=0.4.

Figure 2.7: Data accuracy.

2.8.6/ Further Discussions

In this section, we give further consideration to our proposed techniques. We compare the obtained results while applying the three proposed methods. We give some directions as to which method should be chosen, under which conditions and in which circumstances of the application.

From the energy preserving point of view, the three proposed methods significantly reduce the energy consumption (Fig 2.5). Furthermore, we can observe that the distance methods conserve more energy compared to the K-means with Anova model and the similarity methods. Subsequently, it reduces up to 39% and 64% of the energy of the CH compared to KAB and PFF respectively. Therefore, the Euclidean distance function is the best solution when conserving energy becomes primary. This can be applied when the batteries of sensor nodes reach their lowest levels.

From the data latency point of view, we deduce that the k-means with Anova (KAB) method gives the best result in terms of execution time and data latency, compared to distance and similarity methods. Subsequently, KAB can accelerate the execution time at the CH up to 57 and 26 times compared to the best results obtained using distances and PFF. These results are logical because, searching the groups of redundant sets using KAB has a weak complexity compared to search them in pairs using distances methods or in candidates using PFF. Consequently, when the priority for the application is to deliver data to the sink regardless the aggregation error, the KAB method is more suitable.

From the data accuracy point of view at the CHs, the sets similarity method can save the integrity of the collected data with the minimum loss, e.g. up to 2.81%. On the other hand, the distance method gives better results in terms of data accuracy compared to the KAB method. Hence, if the application does not permit flexibility regarding data accuracy, the similarity functions method is more suitable; else, distance functions and KAB methods can be used as a compromise between energy saving and data accuracy flexibility.

To summarize this section, Table 2.1 presents the characteristics of each method regarding energy consumption, data latency and accuracy, and the complexity of each of the proposed methods.

	Energy consumption conserving	Data latency	Data accuracy	Complexity
Euclidean distance	very good	medium	good	O(n^2)
Cosine distance	good	medium	medium	O(n^2)
KAB	good	very good	low	O(n)
PFF	low	good	very good	O($n \times log(n)$)

Table 2.1: Comparison between distance functions, KAB and PFF techniques.

Remark 6: Other results

To confirm their efficiency, the proposed techniques were also applied on real underwater data collected from the ARGO project [82] and on a real experimentation test-bed. Real Crossbow telosb motes were deployed in our laboratory collecting temperature, humidity and light measures. Furthermore, these techniques were compared to other data aggregation and compression techniques. The obtained results, which are not presented in this document, are very similar to those presented in this chapter which indicate the efficiency of our techniques. For more details please refer to [11, 13, 14].

2.9/ Conclusion

Data aggregation is very essential for WSNs where the huge amounts of data collected by the sensors need to be minimized. In this chapter, we studied three different data aggregation techniques for clustering-based large-scale sensor networks. After eliminating redundant data collected by each sensor, we propose three different data aggregation methods allowing CH to eliminate redundant data sets generated by neighboring sensor nodes. The proposed methods are based respectively on the sets similarity functions, the K-means algorithm applied with one-way Anova model and the distance functions. We have demonstrated through experiments on real data measures the efficiency of the proposed techniques in terms of energy consumption, data latency and accuracy.

3

MULTI-SENSOR DATA FUSION: E-HEALTH APPLICATION

Wireless Sensor Networks (WSNs) are deployed in a zone of interest with the purpose of monitoring the environment and detect events of interest. However, many challenges are addressed in WSNs such as limited energy resources, early detection of important events and fusion of large amount of heterogeneous data in order to take decisions. Data fusion is the process of combining raw data from the various sensor nodes to obtain information of higher quality and make more accurate decisions. In this chapter, we propose two data fusion techniques that uses fuzzy set theory and fuzzy inference system to combine data from multiple sensors at the coordinator level. Although our proposed methods can be used in several monitoring applications, in this chapter we will take e-health and Wireless Body Sensor Networks (WBSNs) as context of application. Our objective is to find the severity level of the patient's health condition based on his/her vital signs scores. The proposed techniques are evaluated on real healthcare data-sets and the results show that they assessed the health condition of different Intensive Care Unit (ICU) patients and overcome existing work in terms of energy consumption and data transmission reduction.

3.1/ INTRODUCTION

WSNs are deployed to collect and process data from the environment in order to have a better understanding of the monitored phenomenon, to detect events and to take the proper decision whenever needed. After data gathering, a fundamental issue to be studied in WSN is data fusion. In this situation, data fusion is the way to process and combine the collected data to help users making accurate decisions and obtaining information of greater quality about the events of interest. Data fusion in WSN is commonly used in different application domains, such as military applications, environmental monitoring, Health monitoring, robotics, etc. [83]. However, the application we consider in this chapter is that of e-health while using wireless body sensor networks (WBSN).

In the paste few years, WBSNs emerged as a low cost solution for healthcare applications. This technology ensures a remote and continuous monitoring of the patient's health condition, therefore reducing healthcare expenditures [84]. Most popular and needed monitoring scenarios include the surveillance of the elderly in nursing homes and in-home monitoring of chronic or acutely ill patients, especially after a surgical intervention. Many applications have been addressed in the literature so far such as gait analysis, monitoring vital signs [85], daily activities [86], fall detection systems and stress evaluation systems [87, 88].

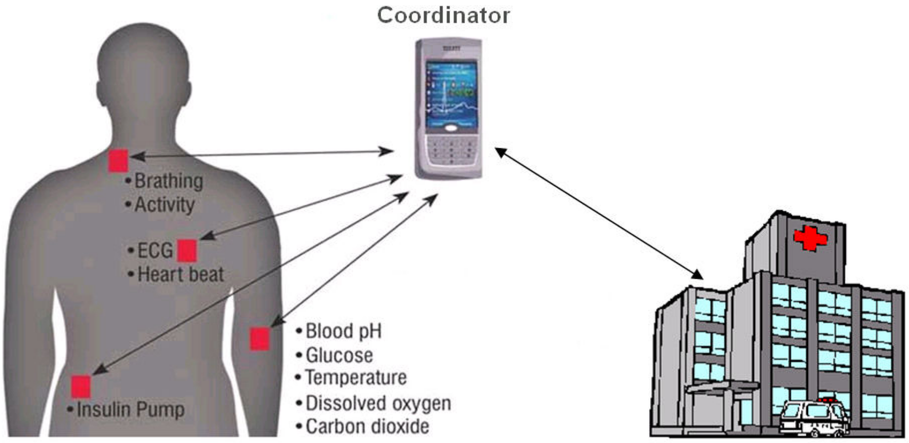


Figure 3.1: WBSN Architecture

The WBSN consists of biosensor nodes and a coordinator (cf. Figure 3.1). First, the nodes are placed on the patient's body and they continuously sense vital signs such as the oxygen saturation, the respiration rate, the skin temperature, etc. [89, 90]. We suppose that each biosensor node only senses one vital sign. Second, the coordinator can be the patient's smartphone, pda or any other portable devices [91]. It receives the collected physiological data in order to perform the multi-sensor data fusion and routinely takes decisions and when emergencies occur. Such emergencies are called critical events since they are triggered when abnormal variations, such as an increase in the heart rate indicating a tachycardia or a decrease in the heart rate indicating a bradycardia, of the vital signs are detected. Moreover, the coordinator alerts the patient when critical events are detected and sends the collected data and the taken decisions to the medical center or any other destination for storage and further analysis [92].

Like traditional WSN several challenges arise in WBSNs. The energy consumed by the biosensor nodes for sensing and transmitting is a highly critical issue, since important physiological variations can be missed out and the data fusion process can be affected if one or more biosensor nodes are dead [93]. Furthermore, the fusion of large amounts of heterogeneous data collected by several biosensor nodes is another challenge in such networks. It enables the coordinator to represent the global situation of the patient and consequently take the corresponding decision.

Several data analysis and processing approaches in WBSNs for anomaly detection, prediction and decision making [94, 95] have been proposed in the literature so far. In the majority of these approaches the data fusion techniques require either offline training, high computation resources or do not take into consideration the energy consumption on the sensor nodes level. In this chapter we study the problem of monitoring and fusing the vital signs of a patient in order to determine the severity of his/her health condition while taking into consideration data reduction for energy consumption requirements. Two data fusion techniques will be presented:

1. The first data fusion scheme uses Fuzzy set theory [96] and a decision matrix. Thus, the coordinator generates appropriate decisions according to the health status of the monitored patient by combining information from various biosensors. The raw data received during consecutive periods are aggregated using fuzzification procedures. Then, the decision having the closest feature values to the aggregated data set is selected from a decision matrix.

2. The second multi-sensor data fusion approach is proposed by defining the input membership functions in terms of the number of vital signs of interest [97]. Fuzzy sets are used to deal with uncertainties and ambiguities and a Fuzzy Inference System (FIS) to map the aggregate score of vital signs to the patient's risk level. We believe that this model is very promising since it is a flexible knowledge-based model and does not require any training. Furthermore, it takes into consideration the uncertainty and the ambiguity that exist in medical data (such as vital signs) that are collected through fuzzy sets and assesses patients' health condition following a human reasoning logic through the fuzzy inference system.

In a second step, a Health Risk Assessment and Decision-Making algorithm (Health-RAD) is proposed. It is implemented on the coordinator of the WBSN that is deployed on the patient's body. Health-RAD employs the proposed multi-sensor data fusion model. It assesses the patient's health condition routinely and each time a critical situation is detected and consequently makes an appropriate decision. We evaluate our approach on real healthcare data-sets and compare it with existing approaches. The results demonstrate the robustness and accuracy of the proposed approach.

The remainder of the chapter is organized as follows. Section 3.2 presents the related work. Section 3.3 presents some background work related to WBSN and vital signs data acquisition. The two multi-sensor data fusion methods are explained respectively in sections 3.4 and 3.5. Then, the health risk assessment and decision-making algorithm is presented in section 3.6. Experimental results are shown and discussed in section 3.7. Finally Section 3.8 concludes this chapter.

3.2/ Related Work

Multi-sensor fusion in WBSN is currently gaining more and more attention since it introduces many advantages in a network that suffers from many limitations such as : data loss, inconsistency and affected sensor samples. It has the potential to reduce uncertainty by increasing the confidence of the collected data and the inferred decisions as well as enhancing the robustness of the healthcare application [98]. Assessing the health condition of a patient suffering from a particular disease or an acutely-ill patient, such as in our scenario, requires a continuous collection of multiple vital signs in order to form a complete view of the patient's situation and perform an accurate health assessment. To this end, multi-sensor fusion is a must to combine and infer heterogeneous data.

Diverse applications based on WBSNs, existing in the literature, propose multi-sensor data fusion techniques such as activity recognition applications, mental health related applications and health monitoring applications.

- Activity recognition: Many researchers have proposed approaches to recognize activities by relying on multi-sensor fusion [99, 100, 101]. For instance, the authors in [102] have studied the sensor fusion impact on activity recognition in order to determine the best combination of sensors and their positions. Feature extraction and selection accompanied by different supervised classification methods are compared.
- Mental health: [85] has proposed a physiological signal classification technique based on multisensor data fusion and case-based reasoning in order to asses the stress level of the individual being monitored. The matching between cases is done using fuzzy logic [88]. A smartphone-based driver safety monitoring system is proposed in [103]. This system is based on data fusion and uses a fuzzy Bayesian network to classify the drowsiness level of the driver.
- Health monitoring: the authors in [104] have designed an algorithm combining sensor selection and information gain allowing a better management of the WBSN. The information gain is defined as the minimum compact set of features required to identify a disease. [105] has proposed a physiological data fusion model for multisensor WHMS called Prognosis. The proposed model generates the prognoses of the patient's health conditions using fuzzy regular language and fuzzy finite-state machine. A framework that performs real-time analysis of physiological data in order to monitor people's health condition is proposed in [106]. The framework determines the severity level of the patient being monitored by computing a global risk. It uses historical data and data mining techniques for model building and performs real-time analysis of the collected vital signs measurements. It has been tested on intensive care unit datasets and the results show that simple K-means has acceptable results and can be used as a clustering algorithm. However, energy consumption due to continuous sensing and transmission was not taken into consideration and the network lifetime was not studied. Furthermore, the health assessment is based on the offline training phase which requires enough medically validated datasets.

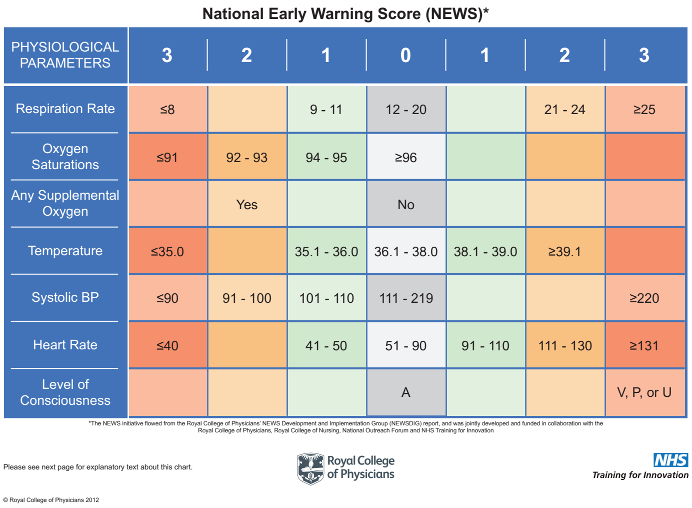
We chose to compare our proposed multi-sensor fusion approach to the approach presented in [106] in terms of accuracy given that the same problem is targeted: patient health assessment. Both approaches ensure a continuous and real-time assessment of the severity level of the patient's health condition based on vital signs monitoring using a WBSN. Furthermore, our complete framework, including the data collection and fusion, is compared to the framework presented in [106] to demonstrate the effect of data reduction on the fusion and the energy consumption in the WBSN.

3.3/ Background

In this section, early warning score systems are presented and the data collection technique which we have adopted in the proposed framework. The former is used by the sensor nodes and the coordinator to assess vital signs. The latter is a previously proposed in Chapter 1 which reduces the amount of sensed and transmitted data to the coordinator, thus extending the network's lifetime.

3.3.1/ Early Warning Score System

An early warning score system (EWS) is a chart used by emergency medical services staff in hospitals to determine the severity level of a specific illness that patients are suffering from or more generally to ascertain their heath status. It is used as a systematic protocol for the measurement and recording of the vital signs. Afterwards, the vital signs are weighed and aggregated in order to allow an early recognition of patients who are subject to an acute illness or those whose health condition is deteriorating [107]. For each vital sign, a normal healthy range is defined. Values outside of this range are allocated a score according to the magnitude of the deviation from the normal range. The score weighing reflects the severity of the physiological disturbance. Since our approach aims at early detecting emergencies, such scoring systems can give the biosensor nodes the ability to locally detect criticalities and only send the important changes in vital signs to the coordinator by computing their scores.



National Early Warning Score (NEWS)*

PHYSIOLOGICAL PARAMETERS	3	2	1	0	1	2	3
Respiration Rate	≤8		9 - 11	12 - 20		21 - 24	≥25
Oxygen Saturations	≤91	92 - 93	94 - 95	≥96			
Any Supplemental Oxygen		Yes		No			
Temperature	≤35.0		35.1 - 36.0	36.1 - 38.0	38.1 - 39.0	≥39.1	
Systolic BP	≤90	91 - 100	101 - 110	111 - 219			≥220
Heart Rate	≤40		41 - 50	51 - 90	91 - 110	111 - 130	≥131
Level of Consciousness				A			V, P, or U

*The NEWS initiative flowed from the Royal College of Physicians' NEWS Development and Implementation Group (NEWSDIG) report, and was jointly developed and funded in collaboration with the Royal College of Physicians, Royal College of Nursing, National Outreach Forum and NHS Training for Innovation

Please see next page for explanatory text about this chart.

Royal College of Physicians

NHS Training for Innovation

© Royal College of Physicians 2012

Figure 3.2: Early Warning System

Figure 3.2 shows the National EWS (NEWS) which has been used in our work. NEWS is standarized and employed in hospitals in the United Kingdom (UK) for the assessment of acute-illness severity [108]. For example, as shown in Figure 3.2, if the respiration rate is between 12 bpm and 20 bpm then the measurement is given a score of 0 indicating that it is in the normal range. However, if the measurement is outside of this range a score of 1, 2 or 3 is given to it according to its level of severity/criticality. For example, if the respiration rate is between 21 bpm and 24 bpm then a score of 2 is given to it. In our work, we have used the measurement ranges defined by NEWS to compute the scores of any of the following vital signs: the respiration rate, oxygen saturation, temperature, systolic blood pressure and heart rate.

3.3.2/ Data Collection

In this section, we adapted the adaptive sampling algorithm presented in Chapter 1 to be executed at the biosensor level. Our goal is to reduce the amount of sensed data by the biosensor node as well as the transmitted measurements to the coordinator [15]. Each biosensor node adapts its sampling rate according to the dynamic evolution of the monitored vital sign and its monitoring importance. This is done using the Fisher Test with One-way ANOVA to study the inter-variances (SF) and the intra-variances (SR) of the collected measurements in m consecutive periods. We define the variable r^0 as the risk level of a vital sign. It represents the monitoring importance given to the vital sign regarding to the patient's health condition such that $r^0 \in [0, 1]$. The greater the value of r^0 is, the more the vital sign is considered critical and the lower its value is, the less the vital sign is considered critical. Having the Fisher Test result F and the risk level r^0, a Quadratic Bezier curve is used as a Behavior Function (BV) to assign the appropriate sampling rate for the following period [109]. On the other hand, our proposed algorithm reduces the transmission by reducing the amount of measurements sent to the coordinator. The biosensor node uses an EWS to detect changes in the state of the monitored vital sign. These changes can indicate a normal state or different levels of criticality. Therefore, the biosensor node sends a measurement each time there is a change in the score indicating an increase or a decrease in the level of criticality.

Algorithm 7: Local Emergency Detection with Adaptive Sampling Algorithm

```
Require: $m$ (1 Round = $m$ periods), $R_{max}$ (maximum sampling rate), $r^0$, $\alpha$
Ensure: $R_t$ (instantaneous sampling rate), $N$ Number of sensed measurements.
  $R_t \leftarrow R_{max}$
  while $Energy > 0$ do
    for each round do
      for each period do
        takes and sends first measurement $r_0$
        gets score $S$ of $r_0$
        takes measurements $r_i$ at Rate $R_t$
        gets score $S_i$ of measure $r_i$
        if $S_i$!=$S$ then
          sends measurement $r_i$
          $S = S_i$
        end if
      end for
      compute $SR$, $SF$ and $F$.
      if $N < m$ then
        $R_t \leftarrow R_{max}$
      else
        find $F_t$ given $\alpha$ such that $F_t = F_\alpha(m-1, N-m)$
        if $F < F_t$ then
          $R_t \leftarrow BV(F, F_t, r^0, R_{max})$
        else
          $R_t \leftarrow R_{max}$
        end if
      end if
    end for
  end while
```

In the following, we assume that all the biosensor nodes run Algorithm 7. All of them have one common period p, at the beginning of which the 1^{st} sensed measurement is sent to the coordinator. During p, a biosensor node senses a measurement at a rate R_t and only sends it if its score is different from the last measurement sent to the coordinator. At the end of each round $R = m \times p$ where m $\in \mathbb{N}^*$, the sampling rate of the biosensor is adapted using the BV function. The latter takes as parameters the maximum sampling rate R_{max} (corresponding to the total of samples in a period), the risk level r^0, the result of the Fisher Test F and the critical F-value F_t as defined by the Fisher Test table for a given Fisher risk α. Noting that R_{max} and r^0 are parameters to be medically judged by the healthcare experts based on the monitoring requirements for a given patient. Further details concerning the energy-efficient data collection technique can be found in [15, 109].

3.4/ Multi-sensor Data fusion model using Decision Matrix and Fuzzy Set Theory

In this section we present a data fusion technique which allows the coordinator to generate appropriate decisions according to the health status of the monitored patient by combining information from various biosensors. The raw data received during consecutive periods are combined using fuzzification procedures. Then, the decision having the closest feature values to the combined data set is selected from a decision matrix.

3.4.1/ Data Fusion and Decision Making

We assume one sensor is activated to monitor one of the features of interest. Assuming there are m features to be monitored, there will be m sensors with the i^{th} sensor observing feature F_i. The m readings are to be aggregated by the data fusion processor to reach a decision concerning the occurrence of an event of interest (e.g. an infraction or an emergency assessment).

We consider that each coordinator fusion processor is provided with a local decision matrix D defined as:

$$D = [D_1, D_2, ..., D_n]$$

where D_k is a vector of score values corresponding to feature values ($[f_{1,k}, f_{2,k}, ..., f_{m,k}]$) and supporting decision d_k.

As an example, let us consider an application where a sensor network is used for the detection if any supplemental oxygen must be given to the patient or not. The set of monitored features may consist of the following:

F_1:	Respiration Rate
F_2:	Oxygen Saturation
F_3:	Heart Rate

The decisions are:

d_1:	supplemental oxygen is needed
d_2:	no supplemental oxygen is needed

A decision matrix can be defined based on any early warning score system. We chose the national early warning score (NEWS) [58] as an example (the elements of the matrix corresponds to the score (NEWS) see figure 3.2):

$$\begin{pmatrix} d_1 & d_2 \\ -- & -- \\ f_{11} & f_{12} \\ f_{21} & f_{22} \\ f_{31} & f_{32} \end{pmatrix} = \begin{pmatrix} d_1 & d_2 \\ -- & -- \\ 1 & 0 \\ 2 & 0 \\ 1 & 1 \end{pmatrix}$$

Where f_{ij} is the score of decision d_j corresponding to feature F_i.

The above example decision matrix indicates that if feature $F1$ (Respiration Rate) is between $9 - 11$, feature $F2$ (Oxygen Saturation) is between $92 - 95$, and $F3$ (Heart Rate) is greater than 91 and less than 110 then decision d_1 is taken (i.e., supplemental oxygen is needed); and if $F1$ (Respiration Rate) is between $12-20$, feature $F2$ (Oxygen Saturation) is greater or equal to 96, and $F3$ (Heart Rate) is greater than 91 and less than 110 then decision d_2 is taken (i.e., no supplemental oxygen is needed).

Given an actual sampling vector S collected by m sensors during one period and composed of m scores corresponding to m readings/features is represented as follows:

$$S = [s_1, s_2, ..., s_m]$$

The coordinator node takes decision d_j so that:

$$\sum_{i=1}^{m}(f_{ij} - s_i)^2 \leqslant \sum_{i=1}^{m}(f_{ik} - s_i)^2 \tag{3.1}$$

for all decisions d_k and for all k going from 1 to n.

Note that $\sum_{i=1}^{m}(f_{ik} - s_i)^2$ is a measure of how close the actual data collected by the sensors is to the feature values expected to support decision d_k. Then, we define the strength of a decision d_k, Str_k, as the inverse of the Cartesian distance as follows:

$$Str_k = \frac{1}{\sum_{i=1}^{m}(f_{ik} - s_i)^2} \tag{3.2}$$

The smaller the distance $\sum_{i=1}^{m}(f_{ik} - s_i)^2$ of the data reading scores S to the feature values expected to supported decision d_k, the stronger the decision.

These decisions can be taken at one period, where only one set of reading is sent to the coordinator. Otherwise, in WBSN the decision is taken based on several reading values corresponding to several number of periods. In the next section, we describe how we use fuzzy sets in order to allow the coordinator to make a decision based on instantaneous and previous data readings.

3.4.2/ Several Data sets fusion

We start this section by a brief overview of fuzzy sets.

3.4.2.1/ Overview on Fuzzy Sets

”Fuzzy sets are sets whose elements have degrees of membership” [59]. A fuzzy set is composed of a set U and a membership function $M : U \rightarrow [0, 1]$. The membership function is a generalization of the characteristic function of an ordinary set.

For each $x \in U$, the value $M(x)$ is called the grade of membership of x in (U, M). It denotes the degree to which the element x is a member of fuzzy set U.

For an ordinary set $U = \{x_1, ..., x_n\}$, we have :

$$M(x) = \begin{cases} 1, & \text{if } x \in U \\ 0, & \text{otherwise} \end{cases}$$

Then x is called not included in the fuzzy set of U if $M(x) = 0$, x is called fully included if $M(x) = 1$, and x is called a fuzzy member if $0 < M(x) < 1$. For fuzzy set, $0 \leqslant M(x) \leqslant 1$. For notational convenience, we do not distinguish between the membership function and the fuzzy set itself. Therefore, the membership function M is the fuzzy set of U. When the set $U = \{x_1, ..., x_n\}$ is finite, we represent fuzzy set M as $M = \{M(x_1)/x_1, ..., M(x_n)/x_n\}$, where, M(x) denotes the degree to which x belongs to M or the confidence of the belief that x belongs to M.

3.4.2.2/ Decision making after several readings

As in the one period readings case, the coordinator uses a decision matrix: $D = [D_1, D_2, ..., D_n]$, where D_l is a score vector for features $[f_{1l}, f_{2l}, ..., f_{ml}]$ of the ideal score values supporting decision d_l. Ideally, if the score values computed by the coordinator (based on its sensor readings) are exactly equal to vector D_l, then the coordinator (i.e., the data fusion processor) should take the decision d_l. However, in the case of several readings sets the computed score values are not the same and will not all match exactly any of the vectors of D. The objective of our decision making process is to select the decision that matches best the computed scores values.

After p periods each sensor took p readings corresponding to p scores of each feature F. This forms a set of scores $S(F) = \{s_1, s_2, ..., s_p\}$ of the feature F. Then, we define the frequency $Freq(s)$ of the score s as the number of the subsequent occurrence of the same score in the same set $S(F)$. For instance, if after 6 readings of the feature F (respiration rate) we obtain $S(F) = \{1, 1, 0, 1, 0, 2\}$ then we have $Freq(0) = 2$, $Freq(1) = 3$ and $Freq(2) = 1$.

Let $S = [\hat{s}_1, \hat{s}_2, ..., \hat{s}_m]$ be a set of fuzzy membership functions computed by the coordinator based on its sensor readings, where $\hat{s}_i$ is the membership function for feature F_i. Note that each $\hat{s}_i$ is a fuzzy membership function computed using the procedure described in the previous section. Using the notation of the previous section:

$$\hat{s}_i = \{M(s_{i1})/s_{i1}, ..., M(s_{ip})/s_{ip}\}$$

In this approach we define the membership function as:

$$M(s_{ik}) = \frac{Freq(s_{ik})}{\sum_{j=1}^{p} Freq(s_{ij})}. \tag{3.3}$$

and the strength of decision d_l as:

$$Str_l = min_{i=1}^{m}(max_{k=1}^{q}(M(s_{ik})e^{-(s_{ik}-f_{il})^2})) \tag{3.4}$$

where $q = |\hat{s}|$ of feature m.

Noting that the min function produces the weakest link among a set of series links and the max function generates the strongest link among a set of parallel links. We use the max function to weed out readings that either have low confidence level or large distance to the ideal value for each feature. Then, we use the min function to represent the strength of a decision with the strength of its weakest feature reading. Other aggregate functions such as the $mean$ can also be used instead of the min and the max functions.

As an example, let us assume the decision matrix below with 2 features F_1 and F_2 and decisions d_1, d_2, and d_3:

$$\begin{pmatrix} d_1 & d_2 & d_3 \\ -- & -- & -- \\ 1 & 3 & 0 \\ 2 & 0 & 1 \end{pmatrix}$$

Assume that the coordinator got the following feature membership values based on its sensor readings and using the fuzzification procedure of the previous section:

$$\hat{s}_1 = \{0.8/1, 0.2/3\}$$
$$\hat{s}_2 = \{0.6/0, 0.4/2\}$$

Intuitively, the readings can support decisions d_1, d_2 and d_3 since d_1 and d_2's values match the readings values but with different levels of confidence and d_3's values are close to the reading values. However it is clear to see that d_1 should be the strongest since its values are the closest to the readings with the highest confidence levels. Applying equation 3.4, we get the strength of each decision:

$$\begin{aligned} Str_1 &= min[max(0.8e^{-(1-1)^2}, 0.2e^{-(3-1)^2}), max(0.6e^{-(0-2)^2}, 0.4e^{-(2-2)^2})] \\ Str_2 &= min[max(0.8e^{-(1-3)^2}, 0.2e^{-(3-3)^2}), max(0.6e^{-(0-0)^2}, 0.4e^{-(2-0)^2})] \\ Str_3 &= min[max(0.8e^{-(1-0)^2}, 0.2e^{-(3-0)^2}), max(0.6e^{-(0-1)^2}, 0.4e^{-(2-1)^2})] \end{aligned}$$

Then we find : $Str_1 = 0.4$, $Str_2 = 0.2$ and $Str_3 = 0.22$. Based on the above results, decision d_1 is the strongest, which intuitively is what we expected, while decision d_2 and d_3 are weaker, with d_2 slightly weaker than d_3.

3.5/ Multi-sensor Data fusion model using Fuzzy Inference System

In this section, we present our second multi-sensor data fusion model based on a Fuzzy Inference System (FIS). This method has as inputs N vital signs collected by N biosensor nodes and as an output the assessment of the patient's health condition which we represent by the patient's risk level. In terms of data processing level of abstraction, the proposed model can be classified under the feature-level fusion category [98].

Figure 3.3 shows the architecture of the proposed model which is composed of the following blocks: the extraction of the up-to-date scores, their aggregation, the mapping to the patient's risk level using a FIS and finally the decision selection. The proposed multi-sensor data fusion approach including all the mentioned blocks (cf. Figure 3.3) is performed by the coordinator of the WBSN. A FIS can determine the patient's risk level using the information it has about how much the patient's health condition is critical. Fuzzy logic is a widely used technique for representing ambiguity in high-level data fusion tasks [110, 111]. Medical data such as vital signs and physiological signals are characterized by uncertainty and ambiguity given that sensor nodes collecting these types of signals are subject to interference, noise and faulty measurements. Moreover, medical data are interpreted in a human reasoning way which enforces the ambiguity presented in such data. Thus, membership functions (MFs) are defined for the input and the output of the FIS and human-language rules are set. In our technique we generalize the membership functions of the input of the FIS in order to make it more flexible and applicable for any number of monitored vital signs.

In the following, we first discuss the extraction of the up-to-date scores which is performed at regular time intervals. Then, we discuss the input of the FIS being the aggregate score and its fuzzification as well as we discuss its output being the patient's risk level. Finally, the whole fuzzy inference system is discussed including the fuzzy rule base as well as the decision-making process.

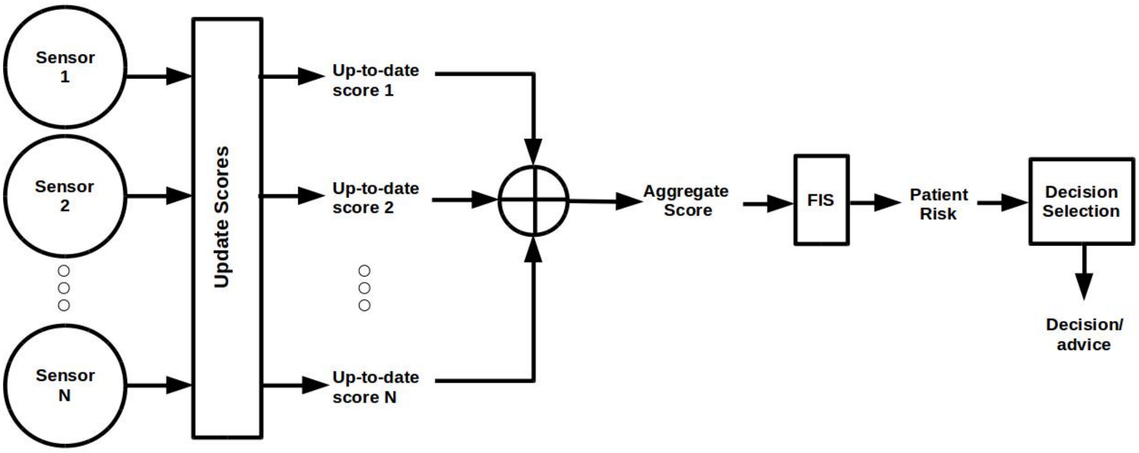


Figure 3.3: Architecture of the Multi-sensor Data Fusion Model

3.5.1/ Up-to-date Score

The biosensor nodes running the data gathering algorithm keep the coordinator updated with changes in vital signs (cf. Algorithm 7). The latter receives several measurements for each vital sign during one round R where $R = m \times p$, $m \in \mathbb{N}^*$. It calculates the up-to-date score s_t for each vital sign at instant t using an EWS as follows:

$$s_t = \frac{s_{t-1} + score_t}{2} \tag{3.5}$$

with $s_0 = score_0$ and where $score_0$ is the score of the first measurement sent during round R, $score_t$ is the vital sign's instantaneous score at time t and s_{t-1} is the score calculated at time $t-1$. Therefore, the instantaneous score $score_t$ and the score s_{t-1}, representing the history of the vital sign, are given equal weights. For example, suppose that biosensor B_1 sends a score of zero at instant $t = 0$. While no other measurement is received during round R, the score s_t of the vital sign is equal to zero. However, if a new score $score_t = 1$ is received at time t, the new s_t would become 0.5 according to equation (3.5). Supposing that no other measurement is received until the end of round R (stable score), if the coordinator updates the vital sign's score s_t each δ_t, then s_t will converge to 1 depending on δ_t and the remaining time until the end of round R such as:

$$\lim_{s_{t-1} \to b} s_t = \lim_{s_{t-1} \to b} \frac{s_{t-1} + b}{2} = b \tag{3.6}$$

where b represents the value of the stable score. Thus, the persistence of a vital sign in the same critical level contributes in the scoring and instantaneous measurements, presenting a deviation, have a lower impact on the scoring.

3.5.2/ Aggregate Score

Health experts and doctors use the aggregate score of the monitored vital signs of a given patient in order to assess his/her health condition. This total score represents the early warning score. It allows them to determine the criticality level of the patient's condition as well as the intervention mode that should be adopted [108]. The aggregate score is used in our approach as an input into the FIS in order to get as an output the patient's risk level. It is calculated as follows:

$$AggScore = \sum_{i=1}^{N} s_i \tag{3.7}$$

where s_i is the up-to-date score (see equation 3.5) of the i^{th} vital sign during a round R and N is the number of monitored vital signs (biosensors).

The analysis and the interpretation of medical data is ambiguous and vary from one subject to another, thus we believe that the assessment of the patient's health condition should be done using fuzzy theory. The input of the FIS is the aggregate score $AggScore$ (see equation 3.7). First, the input is fuzzified using 3 fuzzy membership functions: Low, Medium and High. Then, the process of determining the patient's risk level is executed using a set of fuzzy logic rules.

The aggregate score fuzzy membership functions$f_1(x)$ (Low), $f_2(x)$ (Medium) and $f_3(x)$ (High) are defined as follows:

$$f_1(x) = \begin{cases} 1, & x \leqslant 1 \\ \frac{1}{1-N}x + \frac{N}{N-1}, & 1 \leqslant x \leqslant N \\ 0, & otherwise \end{cases} \tag{3.8}$$

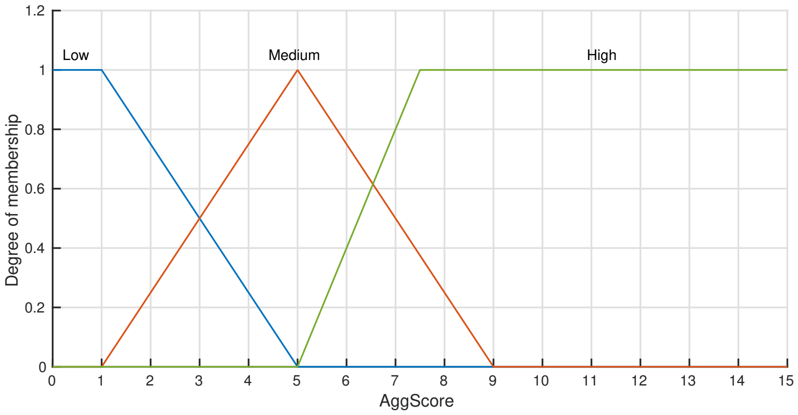


Figure 3.4: Aggregate Score Membership Functions

$$f_2(x) = \begin{cases} \frac{1}{N-1}(x-1), & 1 \leqslant x \leqslant N \\ \frac{1}{1-N}(x+1-2\times N), & N \leqslant x \leqslant 2N-1 \\ 0, & otherwise \end{cases} \tag{3.9}$$

$$f_3(x) = \begin{cases} 2(\frac{x}{N}-1), & N \leqslant x \leqslant \frac{3}{2}N \\ 1, & x \geqslant \frac{3}{2}N \\ 0, & otherwise \end{cases} \tag{3.10}$$

where x represents the aggregate score $AggScore$ and N is the number of monitored vital signs. The definition of these functions was inspired by EWS and the medical analysis carried out by doctors when assessing vital signs and physiological measurements. Figure 3.4 shows the MFs for $N = 5$ vital signs. The aggregate score is Low if $0 < AggScore < 5$, Medium if $1 < AggScore < 9$ and High if $AggScore > 5$.

3.5.3/ Patient Risk Level

As previously mentioned, the objective of this multi-sensor fusion model is to determine the patient's risk level according to the received measurements of the vital signs which are represented by the aggregate score. The patient's risk level r is expressed using a quantitative variable and can range from 0 up to 1. It represents the severity of the patient's health condition. The higher the risk value, the more critical/severe the patient's health condition is. The following fuzzy membership functions are defined for the evaluation of the risk level: Low-Risk, Medium-Risk and High-Risk as shown in Figure 3.5. A patient is at low risk if $0 < r < 0.5$, at medium risk if $0.2 < r < 0.8$ and at high risk if $0.5 < r < 1$.

3.5.4/ Fuzzy Inference System (FIS) and Decision-Making

Figure 3.6 shows the FIS and decision selection blocks of the proposed multi-sensor data fusion model. Having measurements from the N biosensors, the patient's risk level is computed in order to make a decision. The latter is some predictive or corrective advice given to the patient and could be a trigger to a specific action. The input of the FIS is the aggregate score $AggScore$ of the N monitored vital signs (cf. section 3.5.2). Its output is

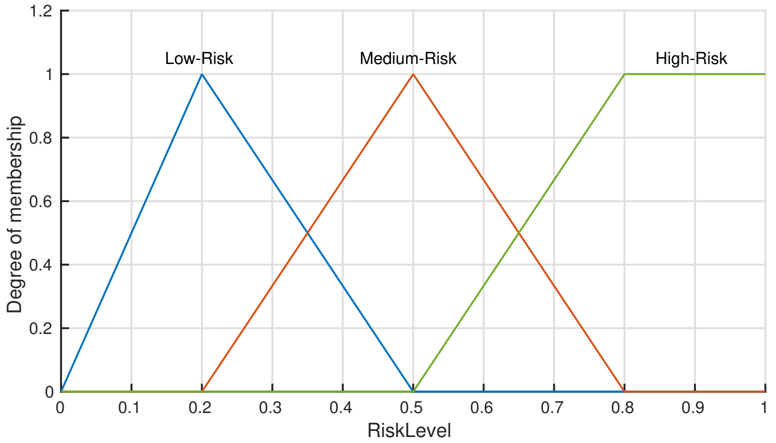


Figure 3.5: Patient Risk Level Membership Functions

the patient's risk level. It uses the fuzzy membership functions described in section 3.5.2 and the fuzzy rule base given by health experts or doctors to map the input to the output.

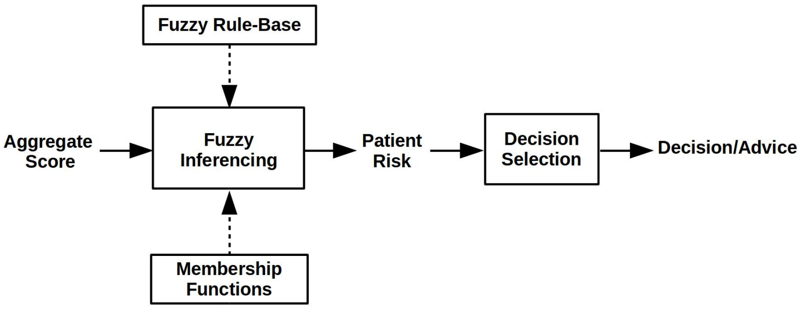


Figure 3.6: Fuzzy Inference System and Decision Selection Blocks

The fuzzy rule base is shown in Table 3.1. For example Rule 1 is: *if the aggregate score is Low then the patient's risk level is Low-Risk*. Finally, the risk level is defuzzified using the centroid method to obtain a crisp patient's risk level r. A decision, some advice or even an action is selected based on the value of r. It is selected from an association table between the patient's risk values and the decisions (c.f. Table 3.2). Such a table is set by healthcare experts. The decisions/advices include for example: rest, take medicine, call the doctor etc. depending on the trigger level. For example if $0 \leqslant r < 0.2$ then decision 1 is taken.

Table 3.1: Fuzzy Rule Base

Rule No.	**Agg Score**	**Patient Risk Level**
1	Low	Low-Risk
2	Medium	Medium-Risk
3	High	High-Risk

Table 3.2: Example of an Association Table between patient risk values and decisions

Decisions	Risk value range
d1	$r < 0.25$
d2	$0.25 \leqslant r < 0.4$
d3	$0.4 \leqslant r < 0.6$
d4	$0.6 \leqslant r < 0.8$
d5	$r \geqslant 0.8$

3.6/ Health Risk Assessment and Decision-Making Algorithm

A Health Risk Assessment and Decision-Making (Health-RAD) algorithm at the coordinator level (cf. Algorithm 8) is proposed based on the data fusion model explained in the previous section. The coordinator receives the measurements sent by different biosensor nodes running $Modified\ LED^*$. Its role is to perform the multisensor data fusion in order to obtain meaningful information about the patient's health condition which is represented by the patient's risk level r. Depending on the value of r, some advice or a decision is given to the patient. The coordinator sends the collected data and the taken decisions to the medical center.

Algorithm 8: Health Risk Assessment and Decision-Making (Health-RAD) algorithm

```
Require: $R$ (monitoring period (round) = $m * p$), $m$ and $p$
   for each round do
      $S_0 = Score_0$
      Reset $AggScore$
      Read received measurement
      Compute $S_i$ for biosensor $B_i$
      Update $S_t$ and $Score_t$
      if $Score_i ! = 0$ then
         Send query to the other biosensors asking for measurements at current time
         Compute $s$ for all biosensors and update $S_t$ and $Score_t$
         Take Decision using fuzzy procedure
         Calculate $AggScore$
         Compute Patient Risk Level
         Select Decision
      end if
   end for
Take global decision at the end of monitoring period (round)
```

The coordinator operates in rounds where round $R = m \times p$ and where p is the common period of all the biosensors at which they are running Algorithm 7 (cf. section 3.3.2) and $m \in \mathbb{N}^*$.

Let $R_0 = (r_1, r_2, r_3, r_4, r_5)$ be the vector of the first measurements received from the 5 biosensors at the beginning of each round. According to the data collection algorithm (Algorithm 7), these measurements are sensed and sent to the coordinator at the beginning of each period p.

Let $Score_0 = (score_1, score_2, score_3, score_4, score_5)$ be the vector of the computed scores corresponding to R_0 and $S_t = (s_{t1}, s_{t2}, s_{t3}, s_{t4}, s_{t5})$ be the vector of the up-to-date scores at instant t.

At the beginning of each round, the coordinator reads R_0, computes $Score_0$ and sets $S_0 = Score_0$. Each time, the coordinator receives a measurement, it identifies the sending biosensor B_i in order to compute $score_i$ using an EWS table and to update $Score_t$ and S_t. Then, it checks whether $score_i$ is different from zero. If this is the case, it detects an emergency and sends a query to the other biosensors in order to get their measurements. After receiving them, the coordinator computes $Score_t$ using the EWS, updates S_t (cf. equation 3.5) and calculates the aggregate score $AggScore$ (cf. equation 3.5). The latter is the input of the proposed FIS. Finally, a decision is selected depending on the patient's risk level given as an output of the FIS. At the end of each round, the $AggScore$ is calculated and a decision is selected based on the result given by the FIS. This decision is a global decision taken routinely, it represents the overall health condition of the patient during one round. Last, S_t is refreshed each δ_t in order to keep track of the patient's condition represented by the scores of his/her vital signs.

3.7/ Performance Evaluation

Experiments are conducted on real medical datasets using a cutom-based Java simulator and Matlab. In order to evaluate the performance of the proposed framework, patient vital signs data-sets are collected from Multiple Intelligent Monitoring in Intensive Care (MIMIC) I, II and III databases of PhysioNet [112]. The default number of monitored vital signs is $N = 5$: heart rate (HR), the respiration rate (RESP), the systolic blood pressure (ABPsys), the blood temperature (BLOODT) and the oxygen saturation (SpO2). Thus, we suppose that 5 biosensors are deployed on the patient's body. In the following, when a different number of vital signs is monitored, the value of N as well as the vital signs of interest will be indicated. The data gathering algorithm (cf. Algorithm 7) is implemented on the biosensor nodes and NEWS (cf. Figure 3.2) is used as a local detection system. The parameters settings for the data gathering algorithm on all biosensors are set as follows:

- Period $p = 100$ sec and a Round is equal to 2 periods.
- Minimum sampling rate $R_{min} = 1$ samples/5 sec and Maximum sampling rate $R_{max} = 1$ sample/2 sec.
- Fisher Risk $\alpha = 0.05$.
- Patient risk $r^0 = 0.9$. Indicating that all vital signs are highly critical and have the same impact on the patient's health.

The parameters settings for Health-RAD algorithm (Algorithm 8, which is implemented on the coordinator, are set as follows:

- $N = 5$ vital signs by default.
- $Round = 100$ sec.
- Update interval $\delta_t = 1$ sec.

The obtained results are compared to the existing approach presented in [106]. The data-sets used in the training phase to build a general intensive care model are taken from MIMIC database and the list is found in [113]. The parameters settings are the following:

- Sampling rate on the sensors: 1 Hz (time granularity of the database 1 measurement/sec).
- Sampling interval on the coordinator: 3 sec.
- Sliding time window size: 10 samples.
- Absolute and Normality thresholds are found in [106].
- k coefficients and h weights for the risk components are found in [113].
- The clustering algorithm: simple K-means.
- The number of risk levels n is set to 4 indicating 4 possible levels : 0, 1, 2 and 3. A higher value of n indicates a higher criticality level.
- The number of clusters for the 3 risk components: $C_{max} = 5$

In the rest of this chapter, we refer to the existing approach [106] that is chosen from the literature as data mining based (dmb) framework.

In the dmb framework, the signal (vital sign) features: offset, slope and distance are used to compute the following risk components: sharp changes, long-term trends and distance from normal behavior (formulas are found in [106]). Then, the health risk associated to signal (vital sign) x at time t is obtained by combining its risk components as follows

$$risk_x(t) = \frac{\sum_i k_{i,x} C(z_i(x))}{\sum_i k_{i,x}} \times \frac{n}{C_{max}}$$

where i ranges from 1 to 3 for the three z_i risk components, $k_{i,x} \in [0,1]$ are weights for the i^{th} component of signal x, C_{max} is the number of discrete levels (the same for every risk component) set during model building and $C(z)$ is the function returning the risk level associated to risk component z. The risk function is normalized to return a value indicating the severity level from 0 to n. Finally, the risk levels of each vital sign are combined together in order to obtain a global risk level for the patient as

$$risk(t) = \max_{x \in X}(risk_x(t))$$

where X designates the monitored vital signs.

The two approaches are compared on the following levels, for different patient records and different number of monitored vital signs:

- Data Reduction
- Energy Consumption
- Vital Signs Assessment
- Health Assessment

Furthermore, our proposed approach is validated against the assessment of a medical expert.

3.7.1/ Data Reduction

The signals of the original dataset of a given patient are shown in Figure 3.7. The dataset is taken from MIMIC II (s01840-3454-10-24-18-46nm record). The signals show the variation of the 5 vital signs of interest over approximately 2 hours, where the sampling rate is set to 1 Hz for all vital signs.

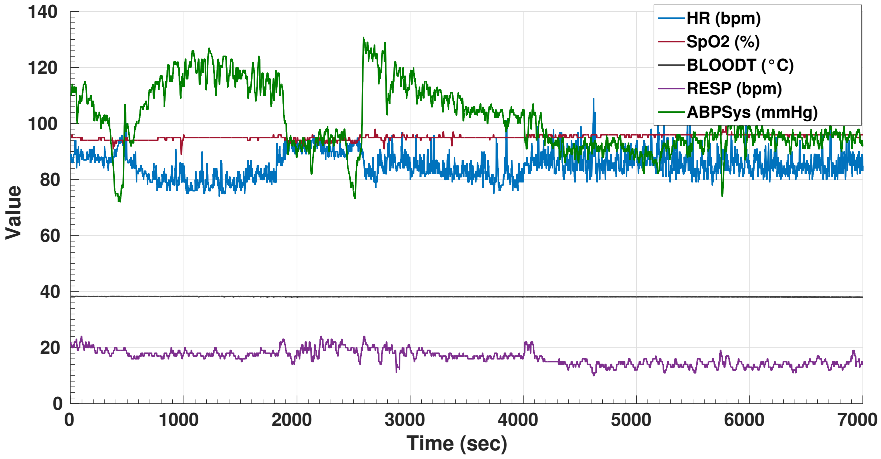


Figure 3.7: Original vital signs signals.

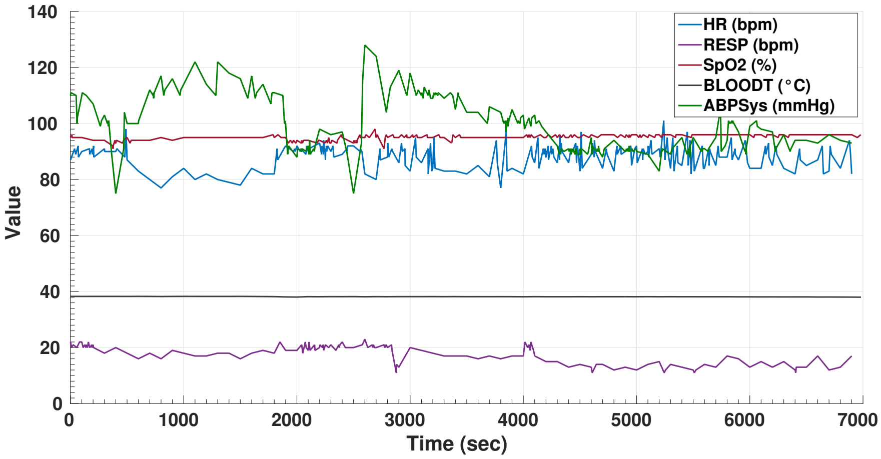


Figure 3.8: Received vital signs signals.

Figure 3.8 shows the signals that are sent to the coordinator over 70 periods after running Algorithm 7, where each signal is sent by a biosensor node sensing the corresponding vital sign. When comparing the original signal of the HR (Figure 3.7), for example, to the sent signal by the HR biosensor (Figure 3.8), it is remarkable to see that the number of small oscillations is considerably reduced while maintaining the general shape and progression of the HR curve over time. This is due to the data gathering algorithm, where only the 1^{st} measurement and changes in the vital sign's score are sent to the coordinator in a period p.

Thus, the amount of redundant data in a period p is reduced and only informative measurements, indicating a decrease or an increase in the vital sign's score, are sent.

Hence, the shape and the progression of the HR curve over time are conserved. An overall data reduction of about 97% is performed compared to the original dataset, while maintaining information about changes in the 5 vital signs' score.

For different patient records and different number of monitored vital signs, Tables 3.3 and 3.4 show the percentages of data reduction performed at the sensing level and the transmitting level in our framework (biosensor nodes running $Modified\ LED^*$) compared to the existing approach [106] in which data are sensed and transmitted each 1 second. The results obtained are over 70 periods. The requests sent by the coordinator running Health-RAD, when critical situations are detected, are taken into consideration in the calculations corresponding to our framework. Missing values in the datasets are ignored and not taken into consideration.

Table 3.3: Percentage of data reduction.

Vital Sign	**Reduction of sensed data (%)**	**Reduction of transmitted data(%)**
HR	63.33	96.91
SpO2	79.58	96.85
BLOODT	64.81	96.93
Resp	72.11	95.56
ABPsys	68.08	95.53

Table 3.4: Comparison of data reduction of four patient records.

Database	**Patient Record**	**Monitored vital signs**	**Reduction of sensed data (%)**	**Reduction of transmitted data(%)**
MIMIC	276n	HR, ABPsys	69.91	88.03
	039n	HR, SpO2, RESP, ABPsys	69.73	92.2
MIMIC II	s01840-3454-10-24-18-46nm	HR, SpO2, RESP, ABPsys, BLOODT	67.87	94.09
	s15480-2803-10-21-19-54nm	HR, SpO2, RESP, ABPsys, BLOODT	69.57	96.36

3.7.2/ Energy Consumption

In this section, we study the energy consumed by the biosensor nodes for sensing and transmitting. The remaining energy after 36 periods in the WBSN in the case of our framework and in the case of the data mining based (dmb) framework are compared. Figure 3.9 shows the results obtained for patient records s01840-3454-10-24-18-46nm (MIMIC II), 039n (MIMIC I), 3000190 and 3100038 (MIMIC III): our framework (in red) vs dmb framework (in blue). We assume that the total initial energy of a sensor node is arbitrarily fixed to 3200 units. The total initial energy in the WBSN is then $N \times 3200$ where N can be equal to 2, 3, 4 or 5. The node consumes 0.04 units for sensing, 0.4 units for transmitting (TX mode) and 0.4 units for receiving (RX mode) [114]. For example, for patient record s01840-3454-10-24-18-46nm, at the end of 36 periods the remaining energy in the WBSN in the case of our framework is about 15010.81 units, however it is only about 8080.0 units in the case of the dmb framework, suggesting that the energy consumption in the WBSN implementing our framework is about 8 times less than the data mining based framework at the end of 36 periods. The number of vital signs of interest N has been varied and the results show that: at the end of one hour, the average energy consumption in the WBSN when applying the proposed approach is approximately 6 times less than the energy consumption in the WBSN when applying the data mining based approach such as the vital signs of interest are the following: HR and RESP (record 300190) and is 16 times less such as the vital signs of interest are the following : HR, RESP and SpO2 (record 3100038) and about 10 times less for record 039n where the vital signs of interest are the HR, REP, SpO2 and ABPSys. Therefore, our approach considerably reduces the energy consumption on the biosensor nodes and extends the WBSN lifetime.

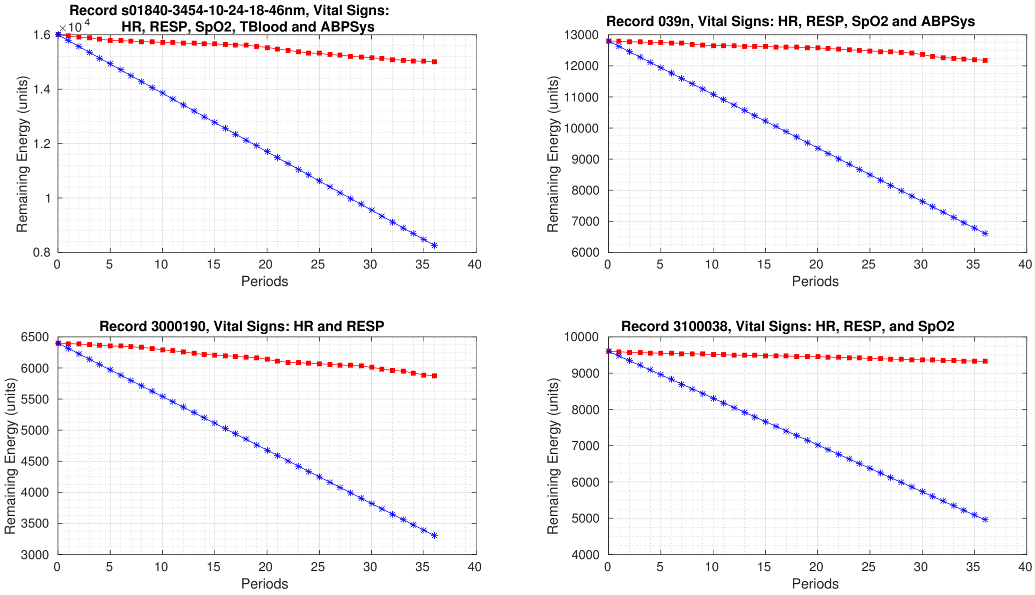


Figure 3.9: Comparison of the remaining energy.

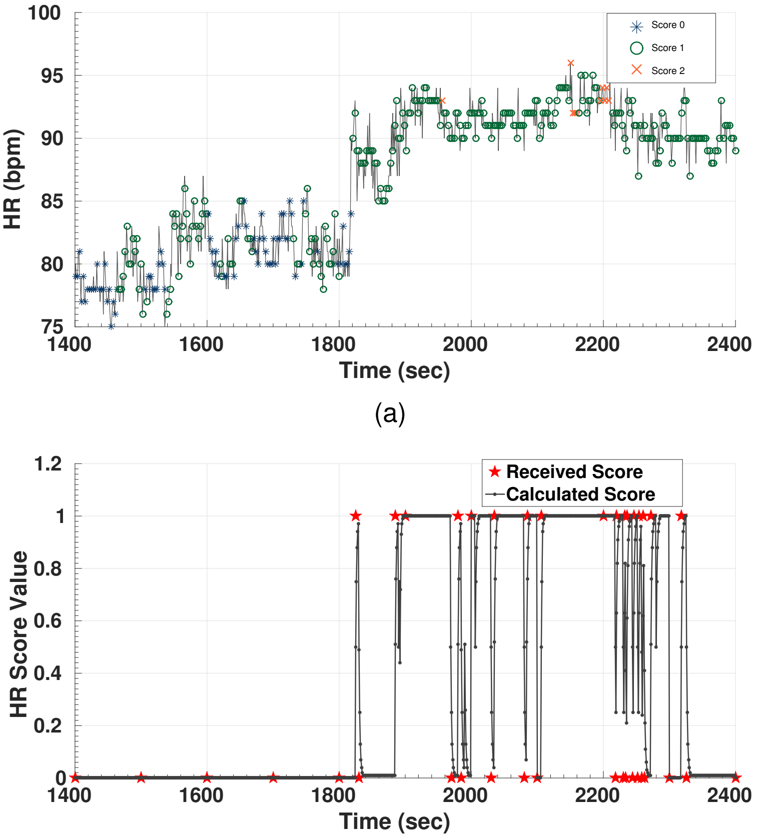
In the following, we compare the results of the two multi-sensor data fusion approaches of the two frameworks. We start by comparing the results obtained at the level of the analysis of the measurements for several vital signs for different patients. Then, we compare the results obtained in the assessment of the patient's health condition (severity level) after performing the data fusion in both frameworks.

3.7.3/ Comparison of the severity level assessment of vital signs

In our approach, Health-RAD regularly updates the scores of the monitored vital signs. In addition, the severity level of a given vital sign is represented by a score between 0 and 3 with score $\in \mathbb{R}$. According to the proposed multi-sensor data fusion model, the score of each vital sign is updated each δ_t and each time a measurement is received from a given biosensor node indicating a change in the status of the vital sign including critical situations. Using equation 3.5, the update of the scores is done while taking into consideration the history and the current score of the vital sign during one round R. As for the dmb framework, the severity level of the vital sign is represented by a risk variable taking values between 0 and $n - 1$, where n is the number of severity levels specified by the user and risk $\in \mathbb{N}$. We set $n = 4$ since the scoring system used in our approach uses four levels ranged between 0 and 3. Figures 3.10 and 3.11 show the assessment of the HR and the SpO2 of patient record s01840-3454-10-24-18-46nm during 1000 sec and 2000 sec respectively. The time intervals were chosen randomnly. On the one hand, Figures 3.10a and 3.11a show the scores assigned to the HR and SpO2 respectively, when applying the data mining based framework which relies on feature extraction and clustering (K-Means) for the online classification. On the other hand, Figure 3.10b and 3.11b show the scores assigned to the same vital signs during the same time interval, but when applying Health-RAD. In Figure 3.10b, the score of the HR is stable and is equal to zero from $t_1 = 1400$ sec until $t_2 = 1800$ sec, indicating that it is normal and not critical. Indeed, according to the measurements of the HR between t_1 and t_2, the values vary between 75 bpm and 87 bpm (cf. 3.10a) which corresponds to the normal range according to NEWS (cf. 3.2). However, Figure 3.10a shows that the score of the HR between t_1 and t_2 vary between 0 and 1 but is, most of the time, equal to 0. Therefore, K-Means has not classified all the HR signal as normal, since at some instants, it was assigned a score of 1. Yet, most of the HR signal between t_1 and t_2 was considered as normal.

After $t_2 = 1800$ sec, Figure 3.10b shows that the calculated score values are between 0 and 1. However, for long time intervals and most of the time, it reaches stability and takes a score of 1. This is due to the stabilization of the received score to 1. When a new score is received, Health-RAD does not affect it automatically to the vital sign, instead it computes a new score based on the last calculated score (history) and the new one received. Since, the fact that a patient has an instantaneous measurement in another score range does not necessarily indicate that his/her health condition is degrading or improving. It is his/her persistence in such conditions which contributes to the risk level. The score of the HR reaches 0 for very short time intervals and this is due to the fast alternation of the HR measurements between score 0 and 1. Hence, our approach assigns to the HR scores between 0 and 1 until stability. Figure 3.10a shows that the HR is assigned most of the time a score of 1, which is compatible to the resuts we obtained in our approach, however K-Means classified it for some instants in a higher risk and assigned it a score of 2. Figures 3.11a and 3.11b show the assessment of the SpO2 during $t_{start} = 2000$ sec and $t_{end} = 4000$ sec. Likewise, both of the approaches assigned alternating scores of 1 and 2 at the beginning. At $t = 2800$ sec, both of them detected a higher level of criticality and assigned a higher score (a score of 3 in the data mining based framework and a score increasing from 2 to 3 in the proposed approach). At $t > 3500$ sec, both of the approaches mostly assigned a score of 1, while the data mining based framework detected some scores of 2.

Likewise, Figure 3.12 shows the assesment of the ABPsys of patient record 267n during 1000 seconds. Both approaches detected high levels of criticality between $t_1 = 2500$ sec and $t_2 = 3000$ sec. Health-RAD assigned to the ABPsys a score up to 3 while the other approach assigned a score of 2.

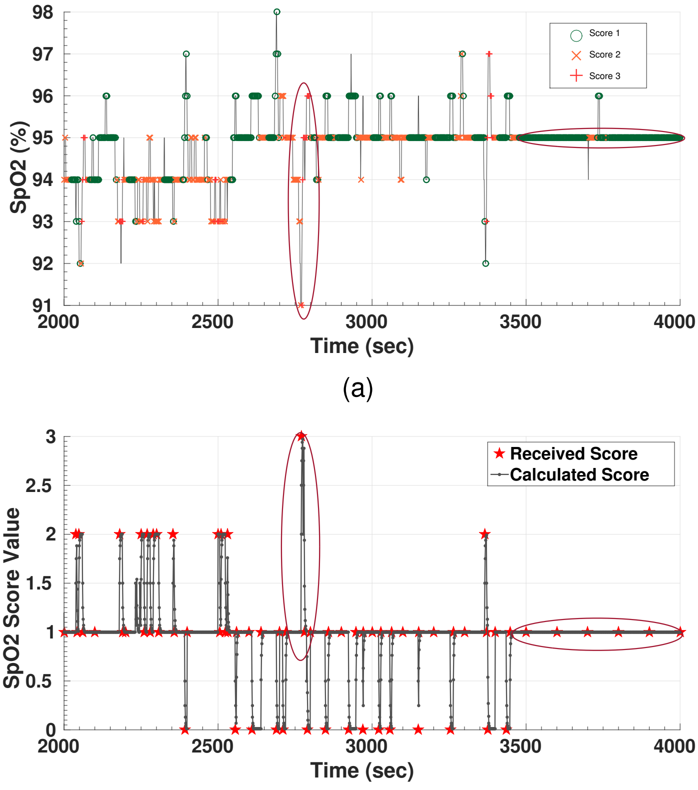


(a)

(b)

Figure 3.10: Severity level assessment of the HR using (a) the dmb framework and (b) our approach

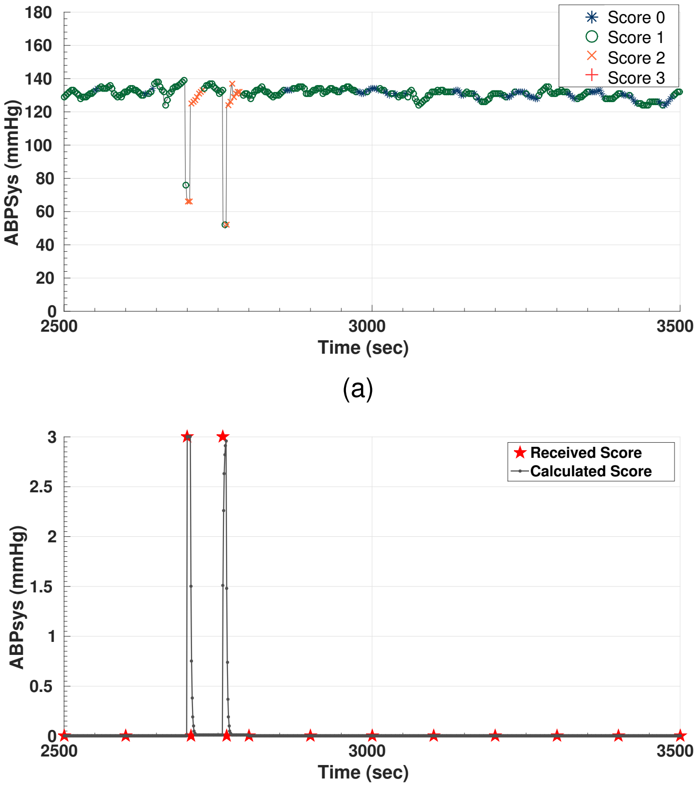
Therefore, the proposed framework analyzed and assessed the vital signs of different patients coherently compared to the dmb approach. However, the proposed approach takes into consideration the limited energy resources requirement in WBSNs. It overcomes the data mining based framework in terms of energy consumption (around 86% less energy consumption) and data reduction (around 70% for sensing and more than 90% for transmission).



(a)

(b)

Figure 3.11: Severity level assessment of the SpO2 using (a) the dmb framework and (b) our approach



(a)

(b)

Figure 3.12: Severity level assessment of the ABPsys using (a) the dmb framework and (b) our approach

3.7.4/ Comparison of the patient health assessment: patient severity level

In this section, we compare the results regarding the patient's health assessment. In both approaches, this is done by performing a multi-sensor data fusion. Figure 3.13 shows the health assessment of the three following patients 3100038, 3000190 and 039n. The first two records are taken from MIMIC III database and the last record is taken from MIMIC I database. For patient 3000190, only the HR and RESP are being monitored, whereas for patient record 3100038 only the HR, RESP and SpO2 are being monitored and for patient 039n the HR, RESP, SpO2 and ABPSys are being monitored (the energy consumption of these records was reported in Section 3.7.2).
In order to compare the risk value of the proposed approach to the global risk of the data mining (dmb) approach, Table 3.5 is used. The average risk per period for each record based on the proposed approach is 0.36 (record 3000190), 0.26 (record 3100038) and 0.53 (record 039n). Thus, the proposed approach has given a global risk of 2 for records 3000190 and 039n and 1 for record 3100038. Similarly, the average global risk per period based on the data mining based approach for record 039n is also 2 and 1 for record 3100038. However, the average global risk per period based on the data mining based approach for record 3000190 is 1. As shown in the plots of record 3100038, both approaches have similarly assessed the patient's health condition over time: the majority of the time the global risk was 1 and alternatively 2. Similarly, as shown in the plots of record 039n, both approaches have in the majority of the time given a global risk of 2 whilst the proposed approach after 2000 sec have alternatively assigned a global risk of 3. For patient record 3000190, the plot of the data mining based approach show that in the majority of the time the global risk was equal to 1 and stable for a longer time compared to when it was equal to 2. Whereas, for the same patient record, the plot of the proposed approach show that a score of 3 was given much more times to the patient's health condition than it was given in the data mining based approach. As a consequence, the average risk per period for record 3000190 was not the same in both approaches.

The results show then that both approaches have detected a critical situation over 1 hour (absence of $risk < 0.2$ and $global\ risk = 0$), that both approaches have similarly assessed the patient's health condition when the vital signs were stable over long periods of time, however the proposed approach reached higher risk values than the data mining based approach when the vital signs presented instability on short time periods and that the data mining based framework is more sensitive to single deviating vital signs.

Table 3.5: Equivalence Table between Risk of our approach and Global Risk of dmb.

Risk	**Global Risk**
$[0, 0.2[$	0
$[0.2, 0.35[$	1
$[0.35, 0.65[$	2
$[0.65, 1]$	3

Tables 3.6 and 3.7 show respectively the average risk per period for 10 records where

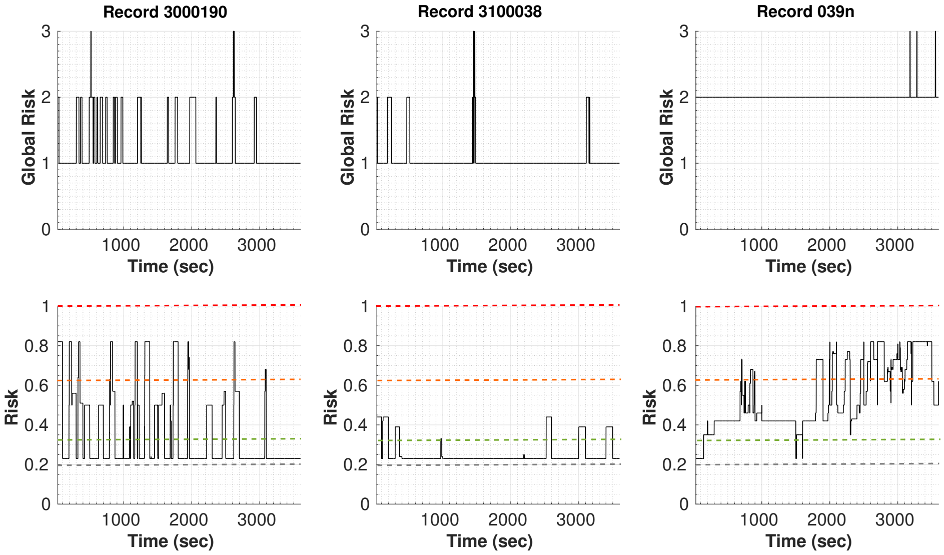


Figure 3.13: Comparison of health assessment during 1 hour between dmb framework (first row) and our proposed approach (second row)

only the HR and RESP are monitored and the average risk per period for 10 other records where only the HR, RESP and SpO2 are monitored based on both approaches. The results show that 50% of the 2 vital signs monitoring records (cf. Table 3.6) have been similarly assessed by both approaches whereas 90% of the 3 vital signs monitoring records (cf. Table 3.7) have been similarly assessed by both approaches. In all the records where the health assessment was different, the proposed approach has given a higher global risk of one class than the data mining based framework (for example patient record 3000190).

Table 3.6: Average Risk per period for vital signs HR and RESP.

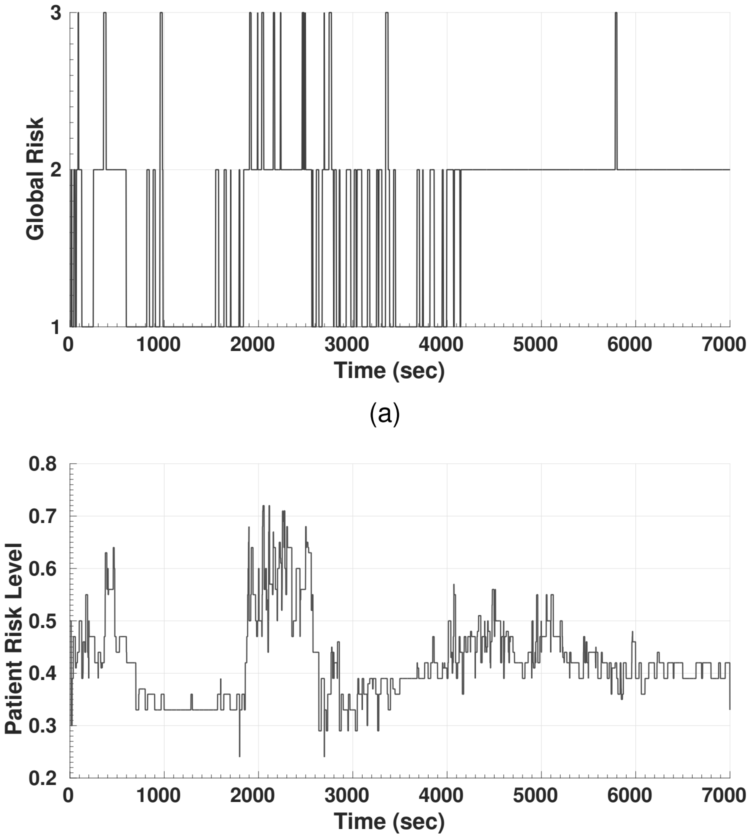
Record	**Average Risk per period**	**Average Global Risk per period**
3000190	0.36	1
3000203	0.33	1
3000598	0.49	2
3000611	0.53	1
3000710	0.27	1
3300295	0.35	1
3300312	0.4	1
3300380	0.23	1
3300430	0.3	1
3300446	0.78	2

Now, a comparison is made based on the default settings of both approaches. In the data mining based framework, the monitored vital signs are the default ones chosen by the authors of [106]: HR, SpO2, ABPdias and ABPsys. In our approach, as per NEWS, the following five vital signs are chosen to perform the patient's health assessment: HR, RESP, ABPsys, BLOODT and SpO2. In the data mining based framework, the patient's

Table 3.7: Average Risk per period for vital signs HR, RESP and SpO2.

Record	**Average Risk per period**	**Average Global Risk per period**
3100038	0.26	1
3100140	0.37	2
3100308	0.23	1
3100331	0.23	1
3100524	0.25	1
3200013	0.33	1
3200059	0.64	2
3200163	0.41	1
3200268	0.26	1
3200359	0.25	1

health condition is represented by a global risk being the maximum of the scores assigned to the monitored vital signs.



(a)

(b)

Figure 3.14: Health assessment using the dmb approach (a) and our approach (b)

This could in some cases trigger false alarms, if it is generated by only one deviating vital sign. This usually occurs when a sensor node is collecting faulty measurements. However, our proposed approach represents the patient's health condition by a patient's risk level. For this purpose, our multi-sensor data fusion model aggregates the scores of all monitored vital signs. Then, it uses the aggregate score as an input into a FIS to generate the patient's risk level. Figure 3.14 shows the results of the health assessment of patient record s01840-3454-10-24-18-46nm during 7000 sec using the data mining based framework and the proposed approach. Clearly, the patient presented high severity levels in the same intervals in both approaches between 2000 sec and 2800 sec and medium severity levels between 4000 sec and 5700 sec and lower ones between 1000 sec and 1500 sec. In our approach, a decision/advice or action is triggered according to the range to which the computed patient risk level belongs.

3.7.5/ Medical domain expert validation

The data collection technique and the EWS based vital sign assessment, used in our framework, have been compared to the classification done by an expert in the medical domain. The comparison focuses on detecting critical events: when the measurements of a given vital sign deviate from the normal range (score $\neq 0$). Table 3.8 shows the results obtained for record s15480-2803-10-21-19-54n for each of the HR, ABPSys and RESP over 28 hours and 46 minutes.

Table 3.8: Accuracy of critical events detection and rate of false positives compared to medical domain expert classification.

	HR	**ABPSys**	**RESP**
Accuracy (%)	93	85	72
False positives (%)	20	15.4	36.3

It shows the accuracy and false positives of the detection of critical events. For each vital sign, we have divided the first 100 000 sec of the record into 100 time frames each of about 1000 sec. If the time frame contains at least one critical event ($score \neq 0$) then it is counted as a positive event, otherwise it is counted as a negative event. The medical expert has classified the 100 time frames based on the knowledge that the record belongs to an ICU patient of a given sexe and age and based on their used vital signs normality thresholds. All of the critical events were detected by our approach for all the vital signs. An average accuracy of about 83% is achieved compared to the expert's classification. However, an average false alarm rate of about 24% is recorded. This is mainly due to narrower normality ranges, which are used in our system, compared to the expert's classification, making it more sensitive to variations. These thresholds can be easily configured depending on the EWS implemented at both the biosensor nodes and coordinator levels.

Remark 7: Application to Stress Evaluation

In order to apply our approach in a real life application, we proposed an energy-efficient stress detection and evaluation framework [115]. A WBSN deployed on the patient's body collects stress-correlated physiological signals. First, the skin conductance (SC) is analyzed. Then, if any stress signs are detected, its level is calculated via our proposed data fusion model and the Fuzzy Inference System (FIS) using the following vital signs: Heart Rate (HR), Respiration Rate (RR) and Systolic Blood Pressure (ABPSys). The results show that the stress evaluation was coherent with the different experimental stages the monitored person has gone through [115].

3.8/ Conclusion

In this chapter, we studied fusion algorithms for data collected from a WSN. These algorithms can be applied in various domains like environmental monitoring, industrial process monitoring, e-health, etc. In this chapter we chose the domain of e-health to test our approach. A two multi-sensor data fusion models has been proposed. The first one is based on a fuzzy data sets theory and the second on a Fuzzy Inference System (FIS). Then, a health risk assessment and decision-making algorithm has been proposed within a complete acute illness monitoring system using a WBSN deployed on the patient's body. A comparison with an existing approach from the literature has been done. The results show that our approach reduces data transmission while preserving the required information. In addition, it reduces the energy consumption due to sensing and transmitting, therefore extending the lifetime of the network of about 10 times over 1 hour of continuous monitoring compared to the other framework proposed in the literature. Furthermore, the assessment of the vital signs and of the global health condition of the patient in both approaches are compatible: risks are detected on time.

4

DATA SURVIVABILITY IN WSN

An important issue related to the WSN is the reliability and the collaboration between sensor nodes in the presence of environmental hazards. High failure rates lead to significant loss of data. Therefore, data survivability is a main challenge of the WSN. It means that the data collected by WSN must be preserved in case of an unreliable network and even in presence of powerful attackers. In this chapter, we develop a compartmental e-Epidemic SIR (Susceptible-Infectious-Recovered) model to save the data in the network and let it survive after attacks. Furthermore, we derive a fully distributed algorithm that supports these models and give the correctness proofs. Numerical simulations are present to support the proposal.

4.1/ INTRODUCTION

The data collected by the nodes in WSN are processed locally and sent to a sink node for further analysis. In a Delay Tolerant Network (DTN), network connectivity is not always available. For example, consider unattended sensor networks where the presence of a sink is sporadic, and it visits the network occasionally to collect data from sensor nodes. This dis-connectivity for a period of time, prevents sensor nodes to offload data in real time and offer greatly increased opportunities for attacks resulting in erasure, modification, or disclosure of sensor-collected data. Moreover, due to the resources limitations of the sensor nodes and in the presence of environmental hasards, failures are more likely to occure. Hence, high failure rates lead to a significant loss of data. For instance, in urban disaster areas, the collected data can identify hazards and save lives whereas nodes failures or attacks lead to data loss. Another example is about critical infrastructure monitoring where data can be lost after a portion of the critical infrastructure suffers a disaster. Thus, data survivability and availability is particularly important in WSN and can not be ignored.

Data survivability helps address the reliability of the data in WSNs. It allows the senor nodes to collaborate and transmit crucial information between them to maximize the amount of monitoring-related data that can survive. Obviously, while addressing data survivability, we need to take into account the unique characteristics of WSNs (the limitation in computing capabilities and energy resources). For that purpose, in this chapter we study and develop a new epidemic-domain inspired approach to model the information survivability in WSN. Somehow, the propagation of the information in a sensor networks could be compared with a disease transmitted by vectors when dealing with public health.

In [116] the authors discussed the spreading nature of biological viruses, leading to infectious diseases in human populations through several epidemic models.

The propagation of the information throughout WSNs can be studied by using epidemiological models for disease propagation. The model we present here is based on the SIR (Susceptible - Infected - Recovered) model. A sensor node is susceptible to a data item when it is online and functioning normally; it can receive the information that must survive. Based on a classical epidemic model, various dynamic models for malicious attacks propagation were proposed [117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127]. The majority of these models were studied for powerful computer networks and based on the fully-connected assumption of the network which is not the case of wireless sensor networks with heavy resources and a very dynamical topology. In this chapter we introduce a thorough analysis of the conditions that can assure data survivability in WSN. We study a new SIR model that considers dynamic topologies and nodes energy constraints. Our novelty in this work is that we study arbitrary dynamic network topologies instead of static networks. We establish a new information propagation model which incorporates the effects of the dynamic WSN topology and its heavy resources. The dynamics of this model are studied, specifically, the level of the attacks and the disappearing of nodes. In a next step, we provide a fully distributed algorithm which supports/covers different epidemic models. The aim of this algorithm is to ensure data survivability in WSN by maintaining a subset of safe nodes in working state while replacing/locking the attacked ones when needed.

The remainder of this chapter is organized as follows: Section 4.2 briefly reviews the state of the art. The SIR model for Data Survivability in WSNs is presented in Section 4.3. Section 4.4 details the proposed epidemic schemes in a comparatively manner. We present in Section 4.5 the design and analysis of the proposed epidemic algorithm and give the proofs. The next section is devoted to numerical simulations. Finally, Section 4.7 concludes this research work.

4.2/ State of the art

In the literature, we can find several mathematical models which illustrate the dynamical behavior of the transmission of biological diseases and/or computer viruses. Based on the Kermack and McKendrick SIR classical epidemic model [128, 129], dynamical models for malicious objects propagation were proposed. Due to the numerous similarities between biological viruses and computer viruses, several approaches and models are proposed to study the spreading and attacking behavior of computer viruses in different phenomena, e.g. virus propagation [130, 131, 132], e-mail propagation schemes [133], virus immunization [134, 135], quarantine [136, 137], vaccination [138], etc. The authors in [139] propose an improved SEI (susceptible-exposed-infected) model to simulate virus propagation. [140] proposes an SEIS-V epidemic model with vertical transmission using vaccination (that is, run of anti-virus software time and again with full efficiency) so that a temporary recovery from the infection of worms can be obtained.

More recently, epidemiological models have been used not only to transmit viruses in computer network but also to ensure the security in wireless sensor networks [141, 142, 143, 144]. The authors in [141, 142] studied the robustness of filtering on nonlinearities in packet losses and sensors. Unattended Wireless Sensor Networks (UWSNs), have been introduced by Di Pietro *et al.* in [145], where adversaries can compromise some sensor nodes and selectively destroy data. In such networks, nodes collect data from the area under consideration, and then they try to upload all the stored data when the sink comes around and the main challenge is data survivability. The epidemiology community has developed the so-called SIR and SIS models [143, 144] of infection. The SIS model (Susceptible - Infected - Susceptible) is suitable for, e.g., the common flu, where nodes may be infected, healed (and susceptible), and infected again. The SIR model (Susceptible- Infected - Recovered) is for example suitable for mumps, where a node, after being infected, becomes recovered (with life-time immunity). SIS, SIR, and SIRS models have been investigated by authors of these research works, in order to derive the parameters that can ensure information to survive. In these articles, the $S(t)$ compartment is constituted by sensors that do not possess the datum at time t, while $I(t)$ is the compartment of sensors that possesses it. Finally, the $R(t)$ compartment is constituted by sensors that have been compromised by the attacker.

The authors in [143, 144] have not taken into account the energy consumption constraints of the nodes. As in WSN, usually the nodes' energy is provided by a battery that can be emptied due to data acquisition, transmission, or simply the functioning cost of keeping nodes alive. On the other hand, the topology of the networks they consider is static, the network's lifetime is unbounded, and nodes cannot die due to empty batteries. Our intention in our study is to provide new epidemic models dedicated to WSN, while taking into account these issues and producing more theoretical results on each model.

4.3/ A SIR model for Data Survivability in WSN

4.3.1/ Introducing the Kermack & McKendrick model

In this section, the SIR model formerly presented in [143, 144] is firstly recalled. Then, consumption hypotheses underlined in this model are precised while theoretical results on the behavior of the compartments of the network are further investigated.

In wireless sensor networks the presence of the sink can be sporadic and disconnectivity is usual to happen. However the duration between two visits of the sink to the network (its absence) can sometimes be considered negligible, in a first approximation, compared to the time required to empty a sensor battery. In such case, the death processes of sensors can be neglected if the aim is to study the immediate consequences of an attack between two visits of the sink. Under such an assumption, the global network can be divided into three compartments, namely the sensors S susceptible to receive the datum of interest (intrusion detection, important information, etc.), the ones that currently store it I, and the recovered sensors R that have been compromised by the attacker: their stored datum has been recovered.

Suppose now that between S and I, the transmission rate is bI, where b is the contact rate, which is the probability of transferring the information in a contact between a susceptible sensor and a sensor having the datum. Indeed, as proven by Di Pietro *et al.*, such a situation occurs when the network is composed of n sensors, and if each sensor forwards the datum with probability $\frac{\alpha}{n}$ [143, 144] (α is the transition rate).

Suppose additionally that the rate to pass between I and R, is c: the attacker is able to individuate the sensors containing the target information, and to destroy each of them with this probability c. Notice that, if the duration of the information survivability is D, then $c = \frac{1}{D}$, as a sensor experiences one recovery in D units of time.

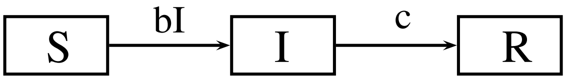


Figure 4.1: SIR model

Under such hypotheses and as stated in [143, 144], the sensors population follows the so-called SIR model of Kermack & McKendrick [128] depicted in Figure 4.1. Remark that the total sensors population is equal to $N = S + I + R = S_0 + I_0 + R_0$, which is a constant: the number of awaken, alive sensors does not evolve. In particular, only two of the three populations of sensors have to be studied.

4.3.2/ First theoretical results

Consider now that $x(t) = \frac{X(t)}{N}$ denotes the fraction of individuals in the compartment X. The SIR model can be expressed by the following set of ordinary non-linear differential equations:

$$\begin{cases} \frac{ds}{dt} = -bis \\ \\ \frac{di}{dt} = bis - ci \\ \\ \frac{dr}{dt} = ci. \end{cases} \tag{4.1}$$

Obviously, the typical time between transmissions is $T_t = b^{-1}$ while the typical time until attack when having the information is equal to $T_e = c^{-1}$. Thus

$$\frac{T_t}{T_e} = \frac{c}{b}$$

is the average number of transmissions between a sensor having the datum and others before it lost this information due to the attacker. Such a statement explain why, in the SIR historical model, the dynamics of the infectious class depends on the *reproduction ratio* defined by

$$R_0 = \frac{b}{c},$$

which corresponds here to the expected number of new informed sensors (so-called “secondary infections”) providing a single sensor with the datum where all sensors are susceptible. Furthermore, direct standard analysis manipulations (variables separation and then integration) lead to the following form for the susceptible sensors compartment: $s(t) = s(0)exp\left(-R_0(r(t) - r(0))\right)$.

As $\frac{di}{dt} = (R_0 s - 1)ci$, if the basic reproduction number satisfies $R_0 > \frac{1}{s(0)}$, there will be an information outbreak with an increasing number of sensors with the datum. In other words, R_0 determines whether or not the information will spread through the network.

All these facts are summarized in the proposition below.

Proposition 1:

Consider a sensor network that aims to monitor a given area, and that has to spread an alert or an information to a sink, whose presence is sporadic. Suppose that an attacker tries to remove the datum in sensors' memory, and that:

1. all sensor activities are negligible, in terms of energy,
2. when a sensor has the datum, it spreads the information to its neighbors with a probability b, until being attacked.

Denote by T_t the typical time between transmissions, T_e the typical time an informed sensor loses its information due to the attacker, and by $s(0)$ the initial fraction of susceptible sensors. So the information will spread through the network if and only if $T_t < s(0)T_e$.

In other words, this proposition states that if the reproduction ratio is greater than one, then an "epidemic" occurs since the prevalence (the infected ratio) increases to a peak and then decreases to zero. Otherwise there is no epidemic since the prevalence decreases to zero.

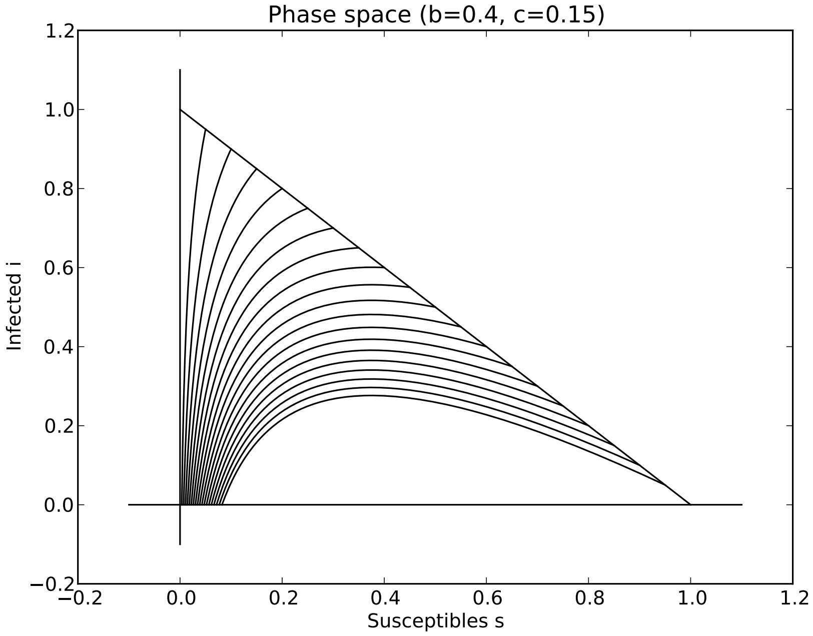


Figure 4.2: Phase space (s,i) with $b = 0.4, c = 0.15$ (SIR model).

It is possible to be more precise in the formulation of Proposition 1, following an approach similar to [146]:

Proposition 2:

The fraction $s(t)$ of sensors susceptible to receive the information is a decreasing function. The limiting value $s(\infty)$ is the unique root in $(0, \frac{T_e}{T_t})$ of the equation

$$1 - r(0) - s(\infty) + \frac{T_e}{T_t}\, ln\left(\frac{s(\infty)}{s(0)}\right).$$

Additionally,

- if $T_t \geqslant s(0)T_e$, then the fractional number $i(t)$ of sensors having the datum decreases to zero as $t \to \infty$,
- else $i(t)$ first increases up to a maximum value equal to $1 - r(0) - \frac{T_e}{T_t}\left(1 + ln\left(\frac{s(0)T_t}{T_e}\right)\right)$ and then decreases to zero as $t \to \infty$, where ln stands for the natural logarithm.

Proof. The triangle $T = \{(s,i) \mid s \geqslant 0, i \geqslant 0, s+i \leqslant 1\}$ is positively invariant, since from the SIR equations, it holds: $s = 0 \Rightarrow s' = 0$, $i = 0 \Rightarrow i' = 0$, and $s + i = 1 \Rightarrow (s+i)' = -ci \leqslant 0$. Furthermore, points on the s axis where $i = 0$ are equilibrium ones, unstable for $s > 1/R_0$ and stable otherwise. s is decreasing and positive due to this invariance and because $\frac{ds}{dt} = -bis$, so an unique limit $s(\infty)$ exists. Similarly, $r'(t) = ci \geqslant 0$ and $r \leqslant 1$ then $r(\infty)$ exists. As $s + i + r = 1$, $i(\infty)$ exists too. To prove that this limit is null, we only remark that if $i(\infty) > 0$, then $r(\infty) = \infty$ (because $r' > \frac{ci(\infty)}{2}$ for sufficiently large t), which is impossible, as $r \leqslant 1$. Finally, the equations of the proposition are derived from $\frac{ds}{di} = \frac{c}{bs} - 1$. □

The phase space of the solutions of the SIR system with given parameters is provided in Figure 4.2 while the evolution of s and i is depicted in Figure 4.3.

The results presented in this section hold for a transition rate between susceptible and informed sensors having the form $F = ai$, which thus represents the force of information. Nonlinear forces of information, or infection, can be investigated too, to model more realistically the information survivability.

4.3.3/ The Recovered compartment

In the previous section, the R compartment was constituted by sensors that have been compromised by the attacker, which will be referred in what follows as situation 1. It is possible to attribute at least two other understandings to this compartment, for an unattended wireless sensor network whose lifetime is dependent on energy consumption and in absence of attacks.

This compartment can be constituted by dead sensors, when considering that the sole action on the energy is the information transmission, and that the unique way to death for a sensor is to have too much transmitted the datum. In other words, in this Situation 2, sensors send information messages to their neighbors until emptying totally

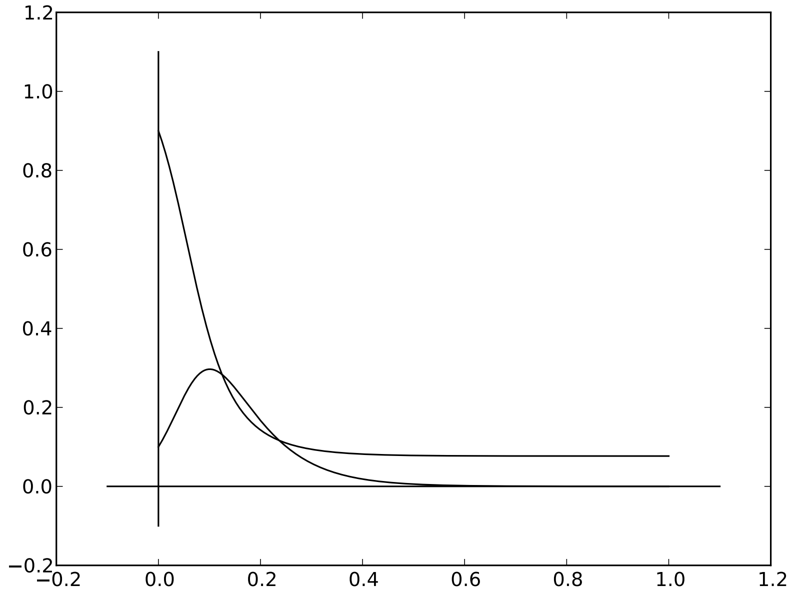


Figure 4.3: Evolution of the fractions s and i of susceptible and having the datum sensors with $b = 0.4, c = 0.15, s(0) = 0.9$, and $i(0) = 0.1$ (SIR model).

their batteries. The sink will receive the information when it will interrogate the network at time t if $I(t) \neq 0$.

A third situation can be considered without any changes in formalization, except redefining the meaning of the R compartment. Indeed, it can be interesting to consider that a sensor is first susceptible to receive an information message for a while, then in a second time it has and transmit the information, before finally entering into the third age of its life, the recovered state in which it will lose its ability to transmit the information. Materials of the previous section tackles too this scenario, when considering the network lifetime sufficiently large compared to information spreading, in order to neglect sensors' death due to energy consumption. The question here is to determine the quantity of informed sensors on large timescales. The three situations of the the R compartment are resumed in table 4.1.

Let us now explain how to extend such a compartmental study for data survivability in wireless sensor networks to well-known SIS models.

4.3.4/ SIS MODELS

Other compartmental divisions of the set of sensors can be investigating, leading for instance to a SIS epidemic model [128]. This latter assumes only two compartments named Susceptible (S) and Infected (I). Transitions between these compartments are represented in Figure 4.4. An individual that is susceptible to a disease becomes infected with a certain probability a, while an infected individual immediately becomes susceptible once (and if) it is cured of an infection (which happens with probability b). Note that a healthy individual can contract a disease only if it is in contact with a sick one. Thus, the

situation 1	Sensors have been compromised by the attackers.
situation 2	Sensors send information messages until emptying totally their batteries.
situation 3	Sensors lose its ability to transmit the information.

Table 4.1: Situations of the R compartment.

evolution of this system is completely described by the following two differential equations (total sensor population: $P = S + I = S_0 + I_0$, which is a constant).

$$\begin{cases} \dfrac{dI}{dt} = aSI - bI \quad I(0) = I_0 \\ \\ \dfrac{dS}{dt} = bI - aSI \quad S(0) = S_0 \end{cases}$$

The SIS model may be treated the same as the SIR model, which has been detailed in this section. For the sake of concision, and as this study does not raise any complication, this model will be left as an exercise, while energy consumption will now be investigated in the next section.

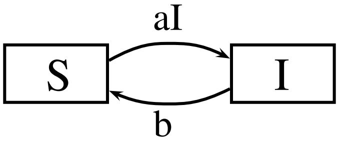
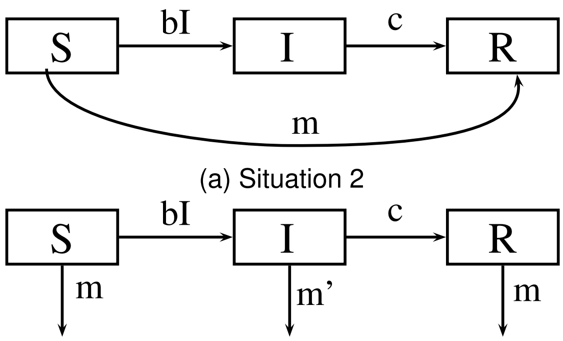


Figure 4.4: SIS model

4.4/ Considering Energy Consumption for Data Survivability in WSNs

In the majority of scenarios and situations, energy consumption and the death of sensors cannot be neglected in WSN. Hence, a ”natural” death rate for all compartments is introduced in this section which leads to generalize the models presented previously.

4.4.1/ A SIR MODEL WITH NATURAL DEATH RATE

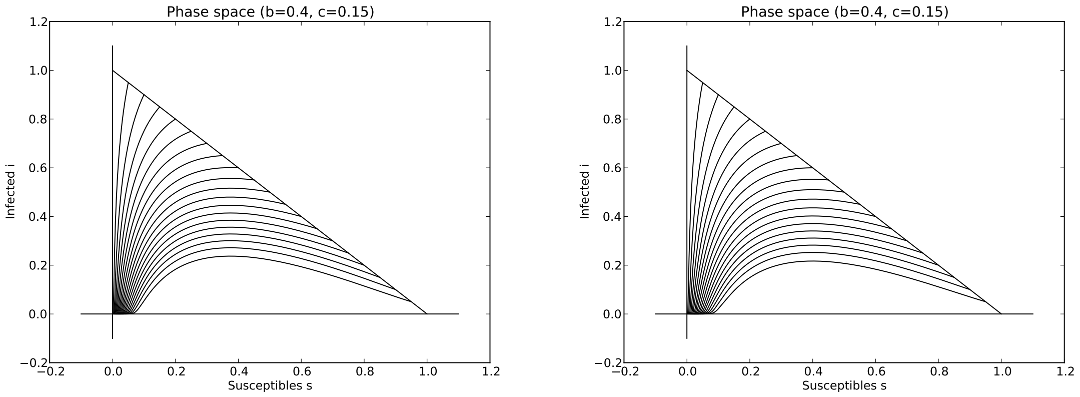


(a) Situation 2

(b) Situation 1 and 3

Figure 4.5: SIR models with natural death rate

The previous section considers that all sensor activities are negligible, in terms of energy and connectivity, except the transmission of information in situations 2 and 3, which is reasonable in a first approximation. It is however possible to refine the SIR model in these two last situations, in order to consider that sensors' energy decreases too in absence of information transmission.



(a) Situation 2

(b) Situation 1 and 3

Figure 4.6: Phase space (s, i) with $b = 0.4, c = 0.15, m = 0.01$, SIR model with natural death rate in Situation 3.

In Situation 2, the R compartment of the SIR model is constituted by dead sensors. This compartment is populated by susceptible nodes that have naturally died (death rate m) without having received the datum and by sensors of the I compartment which die at another rate c supposed to be greater than m, as they have to transfer the datum, an energy-consuming task. This situation is depicted in Figure 4.5a.

In the two other situations investigated in this research work, the R compartment is constituted by living sensors that do not transmit the datum anymore, either because they have been corrupted and thus have lost it (first situation), or because their batteries is preserved (third one). This new situation is closed to the SIR model of Figure 4.1, except that the new network is characterized by a death rate for each sensors compartment (see Figure 4.5b). Notice that the death rate m' of the I compartment is *a priori* different from the one of S and R compartments, as it is reasonable to suppose that the datum transmission implies more energy consumption. However, setting $m' = m$ is possible too.

The SIR model of Equation 4.1 can be adapted as follows for Situation 2:

$$\begin{cases} \dfrac{ds}{dt} = -bis - ms \\[2ex] \dfrac{di}{dt} = bis - ci \\[2ex] \dfrac{dr}{dt} = ci + ms, \end{cases} \tag{4.2}$$

while it has the following form in Situations 1 and 3:

$$\begin{cases} \dfrac{ds}{dt} = -bis - ms \\[2ex] \dfrac{di}{dt} = bis - ci - m'i \\[2ex] \dfrac{dr}{dt} = ci - mr. \end{cases} \tag{4.3}$$

Let us now investigate the long-term behavior of these models. Regarding Situation 2, it is natural to think that, for large timescales, all sensors will take place in the third R compartment of died sensors, as all the batteries are continually emptied (either due to natural consumption or because of the information transmission). This can be easily proven by considering that in an equilibrium point $(s^*, i^*, r^* = 1 - s^* - i^*)$, we have $\frac{ds}{dt} = \frac{di}{dt} = \frac{dr}{dt} = 0$, and so

$$\begin{cases} (bi^* + m)s^* = 0 \\ (bs^* - c)i^* = 0 \\ ci^* + ms^* = 0. \end{cases}$$

As $c > 0, m > 0, i^* \geqslant 0$, and $s^* \geqslant 0$, we can conclude from the third equation above that $s^* = i^* = 0$, and so $r^* = 1$. The Jacobian is equal to

$$J(s,i,r) = \begin{pmatrix} -bi - m & -bs & 0 \\ 0 & bs - c & 0 \\ m & c & 0 \end{pmatrix}$$

and its characteristic polynomial in $(0,0,1)$ is $\lambda(\lambda + c)(\lambda + m)$. The eigenvalues being negative, the equilibrium $(0,0,1)$ is attractive. These results are summarized in the following proposition.

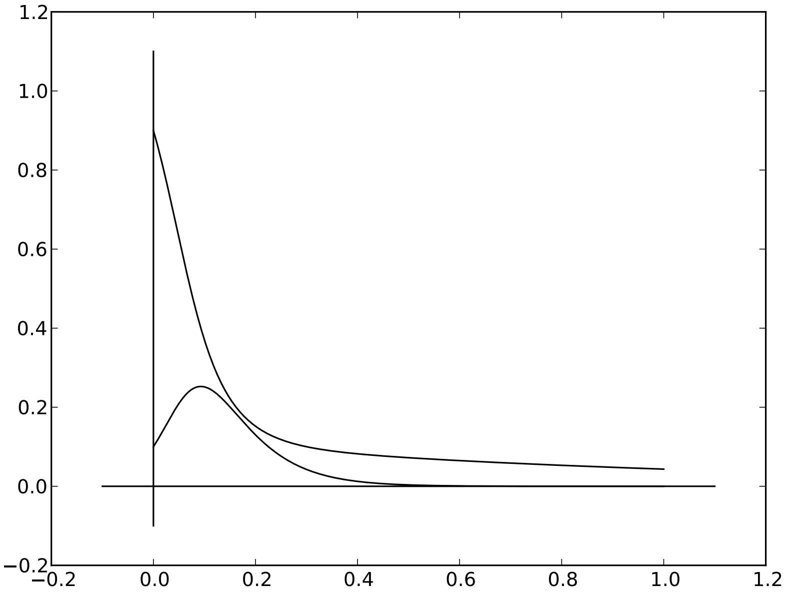


Figure 4.7: Evolution of the fractions s and i of susceptible and having the datum sensors with $b = 0.4, c = 0.15, m = 0.01, s(0) = 0.9$, and $i(0) = 0.1$, SIR model with natural death rate in Situations 1 and 3.

Proposition 3:

Consider an unattended wireless sensor network divided in three sets of sensors, the first category S being susceptible to receive a given datum, the second one I having and transmitting this latter, and the third one R being constituted by dead sensors.
Suppose that the death rate is m for S compartment and c for I's one, and that the transmission rate is bI between S and I. In that situation, for all initial condition and for all positive parameters b, c, and m, the system is convergent to the equilibrium point $(0, 0, 1)$.
In particular, in that situation, the datum cannot survive a long time in the UWSN.

Equation 4.3 can be resolved similarly:

from $bi^*s^* + ms^*$, we deduce that $s^* = 0$ (as $b > 0$, $m > 0$, and $i^* \geqslant 0$). So $bi^*s^* - ci^* - m'i^*$ implies that $i^* = 0$ too. Finally, from the third line, we conclude that $r^* = 0$. Eigenvalues of the characteristic polynomial of the Jacobian in $(0, 0, 0)$ are $-m$ and $-c-m'$, which are negative. So this equilibrium point is attractive too, and a similar proposition than previously can be formulated, with the same conclusion, both for Situations 1 and 3. Phase spaces for the three situations are provided in Figure 4.6 while Fig. 4.7 depicts the evolution of the fractions s and i in Situations 1 and 3.

To put it in a nutshell, to achieve data survivability in WSN, the birth of awaken nodes must be considered, which is the subject of the next subsection.

4.4.2/ A SCHEDULING PROCESS IN DATA SURVIVABILITY

4.4.2.1/ A FIRST NATURAL APPROACH

A first idea to realize a more realistic model of a wireless sensor network is to establish a scheduling process of the sensor nodes, in order to enhance data survivability for a long period of time and enlarge the whole network's lifetime.

Considering the SIR model, such a process leads to the division of each compartment in two parts, corresponding respectively to awaken and to sleeping sensors, as depicted in Figure 4.8.

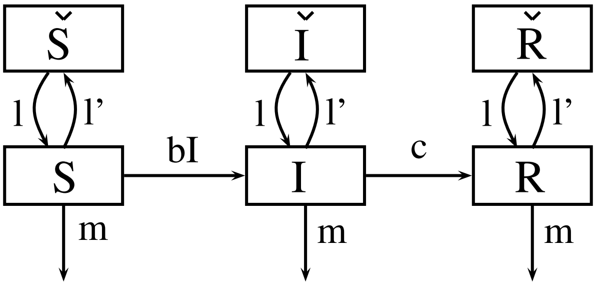


Figure 4.8: SIR model with natural death rate and sleeping nodes

Such a model can be reformulated as follows:

$$\left\{\begin{array}{ll} \dfrac{ds}{dt} = l\check{s} - l's - bis - ms & \dfrac{d\check{s}}{dt} = -l\check{s} + l's \\[2ex] \dfrac{di}{dt} = l\check{i} - l'i + bis - ci - mi & \dfrac{d\check{i}}{dt} = -l\check{i} + l'i \\[2ex] \dfrac{dr}{dt} = l\check{r} - l'r + ci - mr & \dfrac{d\check{r}}{dt} = -l\check{r} + l'r. \end{array}\right. \tag{4.4}$$

The equilibrium point $(s^*, \check{s}^*, i^*, \check{i}^*, r^*, \check{r}^*)$ is searched once again, it satisfies:

$$\left\{\begin{array}{ll} l\check{s}^* - l's^* - bi^*s^* - ms^* = 0 & -l\check{s}^* + l's^* = 0 \\[2ex] l\check{i}^* - l'i^* + bi^*s^* - ci^* - mi^* = 0 & -l\check{i}^* + l'i^* = 0 \\[2ex] l\check{r}^* - l'r^* + ci^* - mr^* = 0 & -l\check{r}^* + l'r^* = 0. \end{array}\right. \tag{4.5}$$

Obviously, $l\check{s}^* = l's^*$, $l\check{i}^* = l'i^*$, and $l\check{r}^* = l'r^*$, and so:

$$\left\{\begin{array}{l} (bi^* + m)s^* = 0 \\[2ex] (bs^* - c - m)i^* = 0 \\[2ex] ci^* - mr^* = 0. \end{array}\right. \tag{4.6}$$

If $s^* \neq 0$, then $bi^* + m = 0$, which is impossible if it is reasonably supposed that each rate is > 0. So $s^* = 0$, which implies that $i^* = 0$, and so $r^* = 0 = \check{r}^* = \check{i}^* = \check{s}^*$.

To sum up, in the unique stable equilibrium point, the number of informed sensors is null, and we face a data loss. This problem is solved in the next section, by considering that nodes never go to sleep.

4.4.2.2/ Achieving data survivability using birth and death rates

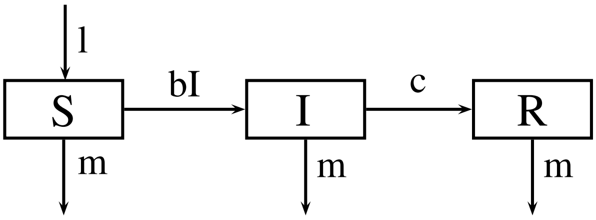


Figure 4.9: SIR model with natural birth and death rates

Consider now a new approach proposed to solve the loss of information in the former scheduling process. In this second approach for scheduling, sensors can only be awaken (we never order them to sleep). It is supposed that a sufficiently large number of sensors are available, and the question is to determine if it is possible to determine the lowest birth rate to achieve data survivability for a long period of time, even in presence of an adversary.

To do so, it is supposed that, at the initial stage, only a small part of the sensors nodes is awakened. New sensors are then awakened periodically during the network's service at a rate l, repopulating by doing so the S compartment (they never go to sleep). Along with this birth rate, a natural death rate m is considered for each of the three kind of sensors, while the R compartment is for corrupted sensors in the original situation 1, as depicted in Figure 4.9. Remark that such a model is compatible with living and awaken nodes that have stopped to transfer the information in Situation 3.

To model such a scenario requires to rewrite the first line of Equation (4.2), leading to the following system:

$$\begin{cases} \dfrac{ds}{dt} = l - bis - ms \\[2ex] \dfrac{di}{dt} = bis - ci - mi \\[2ex] \dfrac{dr}{dt} = ci - mr. \end{cases} \tag{4.7}$$

This updated system is the usual SIR model with vital dynamics, in which we have not supposed the birth and death rates equal. It is possible to show that the problem is well formulated, as the triangle $T = \{(s,i) \mid s \geqslant 0, i \geqslant 0, s+i \leqslant 1\}$ still remains positively invariant.

A study of this system supposes to consider the Poincaré-Bendixon theorem in phase space and the use of Lyapunov functions [146]. It can however be understood by considering what will happen to the information in a long run: will it die out or will it establish itself in the network like an endemic situation in epidemiological models? The long-term behavior of the solutions, which depends largely on the equilibrium points that are time-independent solutions of the system, must be investigated to answer this question. Since these solutions do not depend on time, we have $s'(t) = i'(t) = r'(t) = 0$, which leads to the system:

$$\begin{cases} 0 = l - bis - ms \\ \\ 0 = bis - (c+m)i \\ \\ 0 = ci - mr. \end{cases}$$

$r = \frac{c}{m}i$ from the last equation, and either $i = 0$ or $s = \frac{c+m}{b}$ from the second one. On the one hand, if $i = 0$, then $r = 0$, and $s = \frac{l}{m}$ from the first equation. This leads to the equilibrium solution

$$\left(\frac{l}{m}, 0, 0\right).$$

As the number of sensors having the datum is 0 in this point, it means that if a solution of the system approaches this equilibrium, the fraction i will approach 0, and the datum tends to disappear from the network: an *information-free equilibrium*. Remark that the existence of this equilibrium is independent of the parameters of the system: it always exists.

On the other hand, if $i \neq 0$, then $s = \frac{c+m}{b} \neq 0$ from the second equation, and $\frac{l}{s} = bi + m$ according to the first equation. Substituting s and solving for i, we find

$$i = \frac{bl - m(c+m)}{b(c+m)} = \frac{R_0 l - m}{b},$$

with $R_0 = \frac{bl}{m(c+m)}$, which is a positive number iff $R_0 > 1$.

R_0 is the reproduction number of the information, which tells us how many secondary informed sensors will one informed sensor produces in an entirely susceptible network, as:

- a network which consists of only susceptible nodes in a long run has $\frac{l}{m}$ sensors;
- $c + m$ is the rate at which sensors leave the I compartment. In other words, the average time spent as an informed sensor is $\frac{1}{c+m}$ time units.
- The number of data transmissions per unit of time is given by the incidence rate bIS. If there is only one informed sensor ($I = 1$) and every other sensor is susceptible ($S = \frac{l}{m}$) then the number of transmissions by one “infected” node per unit of time is $\frac{bl}{m}$.

So the number of data transmissions that one informed sensor can achieve during the entire time it is not attacked if all the reminded sensors are susceptible, is $\frac{bl}{m(c+m)}$, that is, R_0.

So if $R_0 > 1$, the number of sensors having the datum is strictly positive in this equilibrium solution: if some other solutions of the system approach this equilibrium as time goes large, the number of sensors having the datum will remain strictly positive, and the information remains in the network and becomes endemic.

These statements are summarized in the following proposition.

Proposition 4:

If either $R_0 \leqslant 1$ or $s(0) = 0$, then any solution $(s(t), i(t))$ is convergent to the equilibrium without information $(1, 0)$.
If $R_0 > 1$ then there are two equilibria: the non attractive information-free equilibrium and the endemic equilibrium. This latter is attractive so that solutions of the ODE system approach it as time goes to infinity: the information remains endemic in the UWSN.

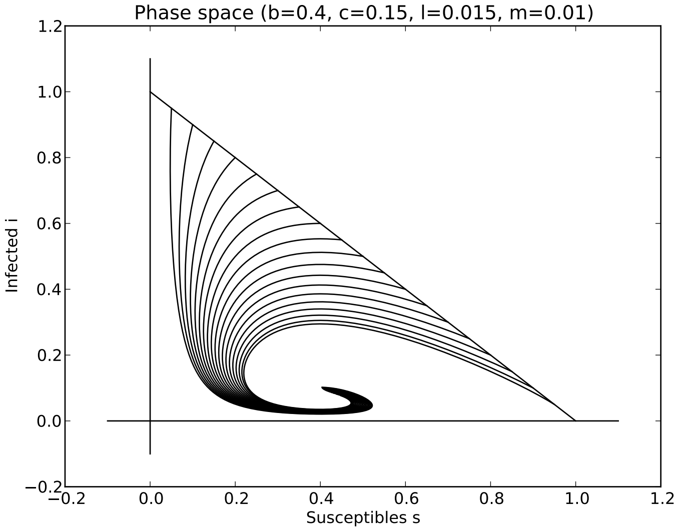


Figure 4.10: Evolution of the fractions s and i of susceptible and having the datum sensors, SIR model with natural birth and death rates ($R_0 = 3.75$).

The attacker desire is to have $R_0 < 1$ to tend to an information-free equilibrium, whereas R_0 must be greater than 1 for the sink to face such attack. If the attacker has the opportunity to observe the network running a certain duration, then he or she can infer the values of parameters b, c, m, and l. Let N be the number of data transmissions by one informed node per time unit, that is, $N = \frac{bl}{m}$. If the attacker is able to detect and infect the informed nodes in a time $\frac{1}{c+m}$ lower than $\frac{1}{N}$, then he or she is sure that $R_0 < 1$: the data will not survive in the network. The sink interest, for its part, is to have $\frac{bl}{m}$ large and $\frac{1}{c+m}$ low, which can be achieved in the following manner:

- increasing the birth rate b,
- increasing the lifetime of sensors to reduce m,
- increasing the data transmission rate b, but m increases when b increases,
- if possible, reducing c by considering countermeasures against data removal.

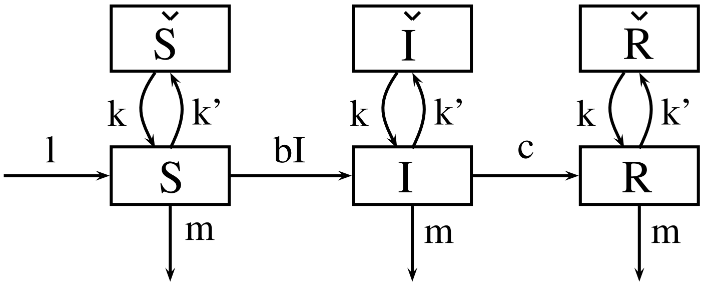


Figure 4.11: Global SIR model with natural birth and death rates, and sleeping nodes

Remark finally that this study is compatible with the situation depicted in Figure 4.11, in which awaken sensor nodes are allowed to go to sleep. Indeed this situation, which has not been detailed in this section to avoid making the text more cumbersome, introduces three new compartments $\check{S}, \check{I}$, and $\check{R}$ as in the previous section. However, as we focused on the future of the information in a long run, we only have to consider equilibrium points that are time-independent solutions of the system. As shown in the previous section, we obtain $\frac{d\check{s}}{dt} = -k\check{s} + k's = 0$, $\frac{d\check{i}}{dt} = -k\check{i} + k'i = 0$, and $\frac{d\check{r}}{dt} = -k\check{r} + k'r$. Consequently, compartments $\check{S}, \check{I}$, and $\check{R}$ disappear in the final global system corresponding to figure 4.11:

$$\begin{cases} \frac{ds}{dt} = l + k\check{s} - k's - bis - ms & \frac{d\check{s}}{dt} = -k\check{s} + k's \\ \\ \frac{di}{dt} = k\check{i} - k'i + bis - ci - mi & \frac{d\check{i}}{dt} = -k\check{i} + k'i \\ \\ \frac{dr}{dt} = +k\check{r} - k'r + ci - mr & \frac{d\check{r}}{dt} = -k\check{r} + k'r, \end{cases} \tag{4.8}$$

and exactly the same Proposition 4 is obtained.

4.5/ Distributed Scheduling Algorithm

In this section, a fully distributed algorithm which supports/covers different epidemic models is presented and theoretically analyzed. Our algorithm seeks to ensure *data survivability* by maintaining a necessary set of safe working nodes and replacing/locking attacked ones when needed.

In the following, we first focus on the legitimate state formulation and next, we present the algorithm which consists in only three rules and give the correctness proofs.

4.5.1/ Problem formalization

Let $G = (V; E)$ the graph modeling the sensor network, with $|V| = n$ and $|E| = m$. We assume sensor node identifiers to be unique. Recall that sensor node identifier is unique if and only if $i.Id \neq j.Id$ holds for each $i, j \in V(i \neq j)$. A sensor node can be in one of these four states: *working*, *probing*, *sleeping* or locked.

We say that a sensor node i is independent if

$$i.state = working \wedge (\forall j \in N_i)\,(j.state = sleeping \vee probing \vee locked)$$

and that i is dominated if

$$(i.state = sleeping \vee probing \vee locked) \wedge (\exists j \in N_i)\,(j.state = working)$$

The legitimate state (let denote it $\mathcal{L}$) of the network is then expressed as follows:

$$\forall i \in V : i.state = working \Rightarrow i.compartment = S \vee I$$

In other words, each working node is either in S or I.

The following notations are also given for the predicates of node i

- $A(i)$: attacked neighbor: $\exists\ j \in N_i, j.compartment = R$
- $W(i)$: working neighbor: $\exists\ j \in N_i, j.state = working$
- $W^*(i)$: working neighbor with lower Id: $\exists\ j \in N_i, j.state = working \wedge i.Id > j.Id$
- $P^*(i)$: probing neighbor with lower Id: $\exists\ j \in N_i, j.state = probing \wedge i.Id > j.Id$

4.5.2/ The algorithm

The proposed algorithm uses the following three rules:

```
r_1:
  if $i.state = probing \wedge W(i)$ then
    if $j.compartment = I$ then
      $i.compartment \leftarrow I$ (*the datum is transferred/replicated to/on $i$*)
    end if
    $i.state \leftarrow sleeping$
  end if

r_2:
  if $i.state = probing \wedge (\neg W(i) \wedge \neg P^*(i) \vee A(i))$ then
    if $A(i)$ then
      $j.state \leftarrow locked$ (*node $j$ remains locked until its healing/recovery*)
    end if
    $i.state \leftarrow working$
  end if

r_3:
  if $i.state = working \wedge W^*(i)$ then
    if $i.compartment = S\ \wedge j.compartment = I$ then
      $i.compartment \leftarrow I$ (*the datum is transferred/replicated to/on $i$*)
    end if
    $i.state \leftarrow sleeping$
  end if
```

4.5.3/ Correctness proofs

Lemma 4:

If a node changes to the *working* state by r_2, then it remains in its state and will never execute a rule again until an eventual attack.

Proof. Let i be a sensor node that executes r_2. According to the preconditions of all rules, node i can execute only rule r_3 in the next round. However, in order to do so, one of its neighbors would have to change into *working* state by r_2. This is impossible as long as node i is in the *working* state. Thus, node i will never execute a rule again. If node i is attacked, it will be locked by r_2 and remains in its state until its healing/recovery. After that, it will join the set of sleeping nodes. □ □

Lemma 5:

If a sensor node is enabled by rule r_2, then each one of its neighbors will execute at most one more rule until their next wakeup/probing, and this rule will be r_1.

Proof. Let i be a node that executes r_2. When node i changes to working state, all its neighbors are either in *sleeping* or *probing* or *locked* state. So we have three possible scenarios: i) neighbors in sleeping state: there is no conflict in this case. ii) neighbors with probing state: those neighbors have a higher Id than i. iii) locked neighbors will remain in their state until their healing/recovery before joining the set of sleeping nodes. □ □

Lemma 6:

Every sensor node is either independent or dominated or locked.

Proof. From the point of view of node i, we have three scenarios:

- if node i is in the $working$ state and is not *independent*, then i may execute rule r_3.
- if node i is in the $sleeping \vee probing$ state and is not *dominated*, then node i may execute rule r_2.
- if node i is in the $locked$ state, then node i will remain in its state until its healing/recovery.

□ □

Lemma 7:

When a node is not locked $\vee$ sleeping, it can make at most 2 moves.

Proof. By Lemma 1 and Lemma 2, each rule can be executed at most once by a node. Hence, the only case a node makes two moves is when it executes r_3 then r_2 with a *working* state. □ □

Theorem 1:

With respect to the legitimate state $\mathcal{L}$ of the network, the proposed algorithm converges within $2n$ moves.

Proof. This follows from Lemma 1 to Lemma 4. □ □

4.6/ Performance Evaluation

This section is dedicated to show some results of the evaluation of the SIR approach through experiments. We will show, using both the mathematical modeling and a basic wireless sensor network, that taking place in the conditions of Proposition 4 is a guarantee to achieve information survivability in WSNs.

4.6.1/ Mathematics-based simulations

In this first illustration, the initial number of susceptible sensors is set to 300 while 3 nodes initially receive the datum. System 4.8 is then discretized and 4 experiments have been conducted, leading twice to the situation $R_0 < 1$, and twice to the opposite situation.

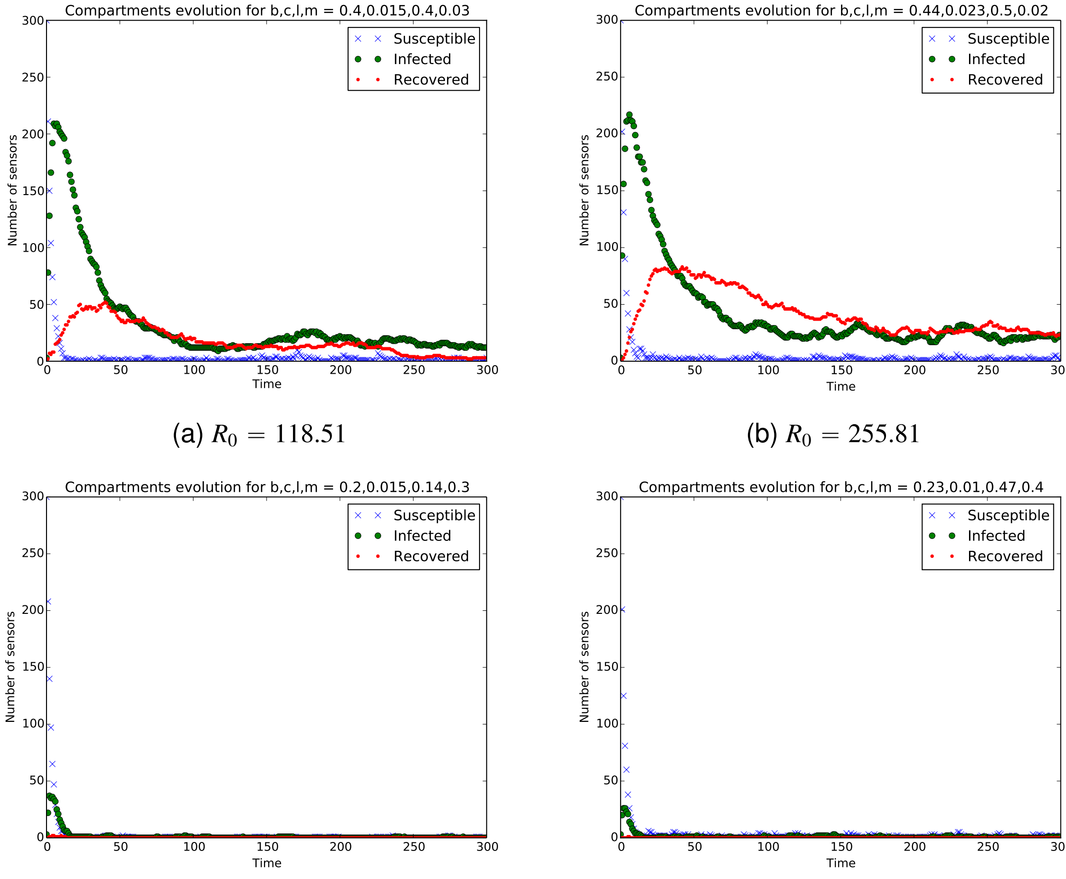
Figure 4.12 shows the obtained results. We can see that the I compartment is never empty when $R_0 > 1$, leading to a data survivability in this SIR model simulation. Conversely, when $R_0 < 1$, the information is obviously lost.

4.6.2/ WSN simulations

In this second set of experiments, we show that the time period of the presence of the information can be extended in a wireless sensor network, and when satisfying Proposition 4.

We have firstly deployed $N = 100$ sensors, all belonging in the susceptible compartment, and with respect to the algorithm detailed in the previous section. In the initial condition, each sensor has a probability of 10% to detect an intrusion (this is the information). At each time unit, an average of lN new sensors are awaken. For each informed sensor and for each of its susceptible neighbor, the data is sent with a probability bI. The death rate of each sensor is set to m (each awaken sensor has the probability m to empty its battery during the considered time unit), while each informed sensor has a probability c to loose the information (to move in the R compartment). The whole network is observed during 60 time units.

We have firstly set $l = 0.017$, $m = 0.0018$, $c = 0.035$, and $b = 0.33$, which leads to $R_0 = 84.69$, and to the situation depicted in Figure 4.13a. In this experiment, * symbols have been used for the susceptible sensors, $\times$ for the informed ones, a circle is for the recovered ones, while the straight line counts the number of dead sensors. A second set of parameters has led to $R_0 = 0.06$, and to the situation described in Figure 4.13b.



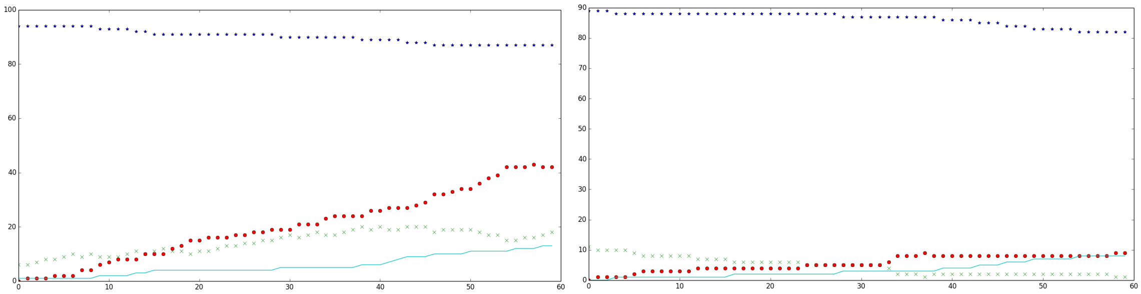
(a) $R_0 = 118.51$

(b) $R_0 = 255.81$

(c) $R_0 = 0.29$

(d) $R_0 = 0.65$

Figure 4.12: Simulation of SIR model with birth and death rates and various R_0



(a) $R_0 = 84.69$

(b) $R_0 = 0.06$

Figure 4.13: Simulation of a wireless sensor network

4.6.3/ 100 Experiments with Random Parameters

We have then launched the previous simulator 100 times with random parameters. At each simulation, probability l is randomly picked in the interval [0,0.2[, m is chosen in

[0,0.01[, c is picked in [0,0.1[, while b is in [0,0.033[, in order to be close to a real situation while having $R_0 < 1$ and $R_0 > 1$ both represented. During these 100 experiments, we have obtained 39 times the situation $R_0 < 1$ with an average of 0.34 (and 61 times the situation $R_0 > 1$, 16.05 of average).

We found an average number of informed sensors equal to 15.50 in the first situation, while it is the double in the second one (33.12 informed sensors in average). In 7 of the 39 simulations with $R_0 < 1$ (17.95%), the number of informed sensors has reached 0, while the information has disappeared 2 times during the 61 other simulations (3.27%). The minimum of informed sensors is attained at the 35-th time unit (in average) in the first situation, while we reach it earlier in the second one (31-th time unit).

To sum up, the information has disappeared in 3.27% of the simulations when $R_0 > 1$, while it has been lost in 17.95% of the cases in the second situation.

4.7/ Conclusion

This paper presented an efficient technique that uses epidemic domain models in the context of data survival in wireless sensors networks. We studied two models (SIR and SIS) that can ensure the survivability of the datum in presence of different types of attacks. We refined the existing SIR model, in order to take into account the WSN constraints, mainly the limited resources and the dynamic topology. In a second step, we proposed and analyzed an efficient distributed algorithm to tackle the problem of data survivability. an energy-efficient fully distributed algorithm that guarantees a necessary subset of sensor nodes to remain non-attacked.

5

Conclusions and Perspectives

This chapter presents the general conclusions of this document and gives some perspectives for future work.

5.1/ Conclusions

The contributions described in this manuscript constitute a summary of my research activities concerning data acquisition and processing in large scale sensor networks. The manuscript is divided into four main chapters detailing some of our contributions related to data management in WSN (collection, aggregation, fusion and survivability).

In the first chapter dedicated to data acquisition and prediction, we presented three contributions related to data collection and transmission reduction in sensor networks. First, we introduced a novel dual prediction mechanism based data reduction algorithm. We have shown that our proposed method is better at reducing the number of transmissions from the node to the sink compared to other existing approaches. Moreover, we took into consideration communication error, links failures, and battery depletion in order to prevent the loss of synchronization between the sink and the node, by applying an appropriate mechanism that identifies and reconstruct missing data. Secondly, we provided an adaptive sampling approach for energy efficient periodic data collection in sensor networks. We studied the sensed data between periods based on the dependence of conditional variance. Then, we proposed a multiple levels activity model that uses behavior functions modeled by modified Bezier curves to define application classes and allow each node to compute its sampling rate. The obtained results show that our approach can be effectively used to increase the sensor network lifetime, while still keeping high quality of the collected data. The third contribution is a combination between the adaptive sampling and dual prediction mechanism techniques. By merging these two techniques together, we were able to reduce radio communication and data sensing at the same time and consequently to preserve a great amount of energy and extend the network lifetime.

The second part of this manuscript is dedicated to our contribution regarding data aggregation in large scale sensor networks. After a preliminary aggregation phase at the node level to reduce local redundancies, three data aggregation techniques were proposed. The first technique proposes a new prefix frequency filtering approach and several optimizations using sets similarity functions to find similar data sets collected from neighbors. It was shown through simulations on real data measurements that this method reduces drastically the redundant sensor measures and outperforms existing prefix filtering approaches. The second technique allows each cluster head to eliminate redundancy from data sent from its members while applying an algorithm based on the k-means algorithm, one way ANOVA model and statistical tests. The experimental results show that this technique largely reduces data redundancy in the network. In the third technique, distance functions (e.g. Euclidean, Cosine, Camberra, Bray-Curtis) are used to search correlation between data sets collected from neighbouring nodes. These three techniques were compared together and to related works. The results show that they outperform existing approaches like data compression or tiny of dag, etc. Furthermore, a discussion is presented outlining under which circumstances each technique is more effective.

The third part of this document focused on our contributions on multi-sensor data fusion. We proposed two data fusion techniques and we considered e-health and body sensor networks as domain of application. We considered a WBSN composed of n biosensors and one coordinator responsible of taking decisions according to the health status of the monitored patient. In our first technique, the aim of the system is to take immediate decisions when an emergency is detected and to monitor continuously the vital signs in order to keep the patient's health stable and under control. Therefore, we have described the data fusion scheme which relies on a decision fusing multiple data sets by using fuzzy procedures. Then, we have proposed a decision-making and monitoring algorithm on the coordinator-level. We have conducted a series of simulations on real medical data recordings to show the effectiveness of our method. The results show that the decisions are taken immediately in an efficient manner, therefore giving the treatment some time to take effect and reducing the amount of unnecessary decisions. In the second technique, a health risk assessment and decision-making algorithm has been proposed within a complete acute illness monitoring system using a WBSN deployed on the patient's body. A generalization of the multi-sensor data fusion model has been proposed in order to make it more flexible and to allow its usage regardless of the number of vital signs being monitored. A comparison with an existing approach from the literature has been done. The results show that our approach reduces data transmission while preserving the required information. Furthermore, the assessment of the vital signs and of the global health condition of the patient in both approaches are compatible: risks are detected on time. These results are validated by a healthcare expert.

The fourth chapter of this manuscript described our contribution on data survivability for unattended wireless sensor networks, where the presence of the sink is sporadic. It allows sensor nodes collaborating and transmitting crucial information between them in order to maximize the amount of monitoring-related data that can survive. We studied two models (SIR and SIS) that can ensure the survivability of the datum in presence of different types of attacks. We showed that our method is well adapted to Unattended WSN scenarios while taking into account the dynamic network topology and the nodes scheduling activities. In a second step, we proposed and analyzed an efficient distributed algorithm to tackle the problem of data survivability.

5.2/ Perspectives

In this section I will highlight short to medium term research directions and perspectives.

5.2.1/ Data reduction from WSN to IoT

Data reduction techniques presented in this document consider WSN with single type of sensor nodes. In the future internet of things, various types of sensor nodes / things will be used. Consequently, various types of data will be sensed and stored. For instance, a health-care application based on wireless body sensor network collect vital signs data and might utilize sensors embedded on smart phones including accelerometer, gyroscope, GPS, microphone, camera, etc. Therefore, sensor nodes will sample multiple types of sensory data (e.g. scalar data, vector data, multimedia data, etc.). The current data reduction work focuses only on a single type of data (e.g. scalar) and rarely consider multi-types sensory data collection and reduction. Thus, the variety of sensory data bring several challenges for efficient data collection. One important issue that should be treated is the correlation among different types of data. Then, based on this correlation, energy efficient cooperative data acquisition techniques need to be designed. For example, stochastic process based methods or time series analysis, can be adopted.

5.2.2/ Collaborative Body Sensor Networks (Interreg Project Response)

In the Interreg Project Response, our objective is to have several heterogeneous sensor networks communicating with each others. In this project, there are two types of networks (active and passive). The first one is a body sensor network (BSN) carried by firemen and the second is a network of temperature and deformation sensors, passive and resistant to high temperatures, fixed in building structures. Such application where multiple individuals' monitoring is required has created a new type of BSN called Collaborative BSN (CBSN), in which data should be gathered and analyzed from multiple bodies rather than a single body to take action accordingly. Even though there are several researches about single BSNs, little studies were found to cover CBSNs. In fact, CBSN is still in its early phases and strong understanding of its architecture and techniques are still lacking. To guarantee a robust and reliable network able to gather and deliver data with high QoS measures, CBSN needs to address several challenges: high mobility, high scalability requirements, coverage and connectivity issues, heterogeneous traffic and irregular traffic pattern, security requirements, etc.

5.2.3/ Deep learning for e-health (ANR Labex Action Project)

The main objective of this project is to study and propose models for recognition and automatic detection of stress to minimize anxiety disorders that have a direct influence on the whole society. Our challenge is to identify stress in different individuals and in different contexts by coupling two recent and promising technologies namely body sensor networks and deep learning. This project consists of two main stages, data collection, and machine learning. The idea is to explore deep learning techniques such as convolutional networks or long short-term memory [LSTM], initially focusing on a small number of classes. We will start by studying a small number of classes, for example: unstressed, light stress and high stress. If we see that it is possible to study finer classes or that it is possible to detect other phenomena, we will increase the number of classes. It is important to have a large enough number of individuals to increase the robustness.

5.2.4/ Privacy applied to e-health

These days, we progressively find ourselves surrounded by smart cyber-physical systems that silently track our activities and collect information about us. Examples include smart homes and cities, remote patient monitoring, WSN, IoT. While such systems may ease our lives, they raise major privacy concerns for their users, as collected data is often sensitive, e.g. vital signs, location. In collaboration with LIRIS laboratory from University of Lyon1, our objective is to address these concerns by proposing a solution that enables the users to play a central role in protecting their privacy. We propose models, techniques and tools to help users, before sharing their data with data consumers, to 1) identify the privacy risks involved in that sharing; 2) assess the value of data, based on identified risks, and compare it to the benefits generated by the sharing; 3) control the data release by applying data modification techniques to implement taken sharing decisions. We intend to apply this solution to the healthcare domain to protect the privacy of patients in smart healthcare environments.

5.2.5/ Modular Robots

An important research axis that I wish to develop in my perspectives is the programmable matters and modular robots. Recently, nano-robots can be used in several types of applications while covering a wide range of tasks. For example, they are used for cleaning arteries, autonomous space exploration, urban search and rescue, Educational purposes, etc. Nowadays, fixed-body robots can perform several tasks in an accurate manner. However, they are not adaptable and flexible to unpredicted environments. The idea of modular robots is to create self-organizing machines that adapt themselves according to the surrounding environment and unexpected events. Employing these new systems generates a large number of challenging problems in design, optimization, and planning. One of my main perspectives is to study these systems and propose proper solution approaches for these challenges.

List of Publications

Articles in refereed international journals

1. Carol Habib, Abdallah Makhoul, Rony Darazi and Raphaël Couturier, ”Health Risk Assessment and Decision-Making for Patient Monitoring and Decision-support using Wireless Body Sensor Networks”, Information Fusion, vol. 47, pp. 10-22, 2019.

2. Moustafa Harb Hassan and Makhoul Abdallah, ”Energy Efficient Sensor Data Collection Approach for Industrial Process Monitoring”, IEEE Transactions on Industrial Informatics, vol. 14, 2, pp. 661 - 672, feb. 2018

3. Gaby Bou Tayeh , Abdallah Makhoul, David Laiymani, Jacques Demerjian, ”A Distributed Real-Time Data Prediction and Adaptive Sensing Approach for Wireless Sensor Networks”, Pervasive and mobile computing, vol 49, pp. 62-75, 2018

4. Christian Salim, Abdallah Makhoul, Rony Darazi and Raphaël Couturier, ”Similarity Based Image Selection with Frame Rate Adaptation and Local Event Detection in Wireless Video Sensor Networks”, Multimedia Tools and Applications, vol. A venir, pp. A venir, 2018

5. Moustafa Harb Hassan and Makhoul Abdallah, ”Energy-efficient scheduling strategies for minimizing big data collection in cluster-based sensor networks”, Peer-to-Peer Networking and Applications, vol. A venir, pp. A venir, 2018

6. Farhat Ahmad, Guyeux Christophe, Makhoul Abdallah, Jaber Ali and Tawil Rami, ”On the coverage effects in wireless sensor networks based prognostic and health management”, International Journal of Sensor Networks (IJSN), vol. A venir, pp. A venir, 2018

7. Makhoul Abdallah, Jaber Ali and Tawbi Samar, ”Energy Efficient Data Collection in Periodic Sensor Networks Using Spatio-Temporal Node Correlation”, International Journal of Sensor Networks (IJSN), vol. A venir, pp. A venir, 2018

8. Farhat Ahmad, Guyeux Christophe, Makhoul Abdallah, Jaber Ali, Tawil Rami and Hijazi Abbas, ”Impacts of wireless sensor networks strategies and topologies on prognostics and health management ”, Journal of Intelligent Manufacturing, vol. A venir, pp. A venir, nov. 2017

9. Moustafa Harb Hassan, Makhoul Abdallah, Laiymani David and Jaber Ali, ”A Distance-based Data Aggregation Technique for Periodic Sensor Networks”, ACM Transactions on Sensor Networks, vol. 13, 4, pp. 32 (40 pages), sep. 2017

10. Moustafa Harb Hassan, Makhoul Abdallah, Couturier Raphael and Tawbi Samar, ”Comparison of Different Data Aggregation Techniques in Distributed Sensor Networks”, IEEE Access, vol. 5, pp. 4250 - 4263, mar. 2017

11. Makhoul Abdallah and Moustafa Harb Hassan, ”Data Reduction in Sensor Networks: Performance Evaluation in a Real Environment”, IEEE Embedded Systems Letters, vol. 9, 4, pp. 101 -104, dec. 2017

12. Moustafa Harb Hassan, Makhoul Abdallah, Jaber Ali, Tawil Rami and Bazzi Oussama, ”Adaptive Data Collection Approach based on Sets Similarity Function for Saving Energy in Periodic Sensor Networks”, International Journal of Information Technology and Management (IJITM), vol. 15, 4, pp. 346 - 363, 2016

13. Habib Carol, Makhoul Abdallah, Darazi Rony and Salim Christian, ”Self-Adaptive Data Collection and Fusion for Health Monitoring Based on Body Sensor Networks”, IEEE Transactions on Industrial Informatics, vol. 12, 6, pp. 2342 - 2352, dec. 2016

14. Makhoul Abdallah, Guyeux Christophe, Hakem Mourad and Bahi Jacques, ”Using an Epidemiological Approach to Maximize Data Survival in the Internet of Things”, ACM Transactions on Internet Technology (TOIT), vol. 16, 1, pp. 5 (15 pages), feb. 2016

15. Makhoul Abdallah, Moustafa Harb Hassan and Laiymani David, ”Residual Energy-based Adaptive Data Collection Approach for Periodic Sensor Networks”, Ad Hoc Networks, vol. 35, *, pp. 149–160, dec. 2015

16. Moustafa Harb Hassan, Makhoul Abdallah and Couturier Raphael, ”An Enhanced K-means and ANOVA-based Clustering Approach for Similarity Aggregation in Underwater Wireless Sensor Networks”, IEEE Sensors Journal, vol. 15, 10, pp. 5483 - 5493, oct. 2015

17. Makhoul Abdallah, Laiymani David, Moustafa Harb Hassan and Bahi Jacques, ”An Adaptive Scheme for Data Collection and Aggregation in Periodic Sensor Networks”, International Journal of Sensor Networks (IJSN), vol. 18, 1/2, pp. 62 - 74, jun. 2015

18. Guyeux Christophe, Makhoul Abdallah, Atoui Ibrahim, Tawbi Samar and Bahi Jacques, ”A Complete Security Framework for Wireless Sensor Networks: Theory and Practice”, International Journal of Information Technology and Web Engineering (IJITWE), vol. 10, 1, pp. 47 - 74, 2015

19. Moustafa Harb Hassan, Makhoul Abdallah, Laiymani David, Jaber Ali and Bazzi Oussama, ”An Analysis of Variance-based Methods for Data Aggregation in Periodic Sensor Networks” in ”Transactions on Large-Scale Data- and Knowledge-Centered Systems XXII” edited by Hameurlain, Abdelkader and Roland Wagner , Josef Küng, Springer, vol. 9430, Lecture Notes in Computer Science (LNCS), pp. 165 - 183, jul. 2015

20. Bahi Jacques, Makhoul Abdallah and Medlej Maguy, ”A Two Tiers Data Aggregation Scheme for Periodic Sensor Networks”, Ad Hoc & Sensor Wireless Networks, vol. 21, 1-2, pp. 77-100, 2014

21. Bahi Jacques, Guyeux Christophe and Makhoul Abdallah, ”Two Security Layers for Hierarchical Data Aggregation in Sensor Networks”, International Journal of Autonomous and Adaptive Communications Systems (IJAACS), vol. 7, 3, pp. 239 - 270, 2014

22. Bahi Jacques, Guyeux Christophe, Hakem Mourad and Makhoul Abdallah, ”Epidemiological Approach for Data Survivability in Unattended Wireless Sensor Networks”, Journal of Network and Computer Applications, vol. 46, pp. 374 - 383, nov. 2014

23. Moustafa Harb Hassan, Makhoul Abdallah, Tawil Rami and Jaber Ali, ”Energy-Efficient Data Aggregation and Transfer in Periodic Sensor Networks”, IET Wireless Sensor Systems, vol. 4, 4, pp. 149 - 158, dec. 2014

24. Bahi Jacques, Guyeux Christophe, Makhoul Abdallah and Pham Congduc, ”Low Cost Monitoring and Intruders Detection using Wireless Video Sensor Networks”, International Journal of Distributed Sensor Networks, vol. 2012, pp. ID 929542 (11 pages), nov. 2012

25. Bahi Jacques, Makhoul Abdallah and Medlej Maguy, ”Energy Efficient in-Sensor Data Cleaning for Mining Frequent Itemsets”, Sensors and Transducers, vol. 14, 2, pp. 64–78, mar. 2012

26. Pham Congduc, Makhoul Abdallah and Saadi Rachid, ”Risk-based Adaptive Scheduling in Randomly Deployed Video Sensor Networks for Critical Surveillance Applications”, Journal of Network and Computer Applications, vol. 34, 2, pp. 783 - 795, mar. 2011

27. Makhoul Abdallah, Bahi Jacques and Mostefaoui Ahmed, ”Localization and Coverage for High Density Sensor Networks.”, Computer Communications, vol. 31, 4, pp. 770 - 781, mar. 2008

28. Makhoul Abdallah, Bahi Jacques and Mostefaoui Ahmed, ”Hilbert mobile beacon for localization and coverage in sensor networks”, International Journal of Systems Science, vol. 39, 11, pp. 1081 - 1094, nov. 2008

Articles in international conferences

1. Christian Salim, Amani Srour, Rony Darazi, Abdallah Makhoul and Raphaël Couturier. ”Enhanced S-MAC Protocol for Early Reaction and Detection in Wireless Video Sensor Networks”, the 17th IEEE International Symposium On Parallel And Distributed Computing (ISPDC 2018)

2. Bou Tayeh Gaby, Makhoul Abdallah, Demerjian Jacques and Laiymani David, ”A new autonomous data transmission reduction method for wireless sensors networks” in ”2018 IEEE Middle East & North Africa COMMunications Conference (MENACOMM 2018)”, Jounieh, Lebanon, pp. (6 pages), apr. 2018

3. Azar Joseph, Makhoul Abdallah, Darazi Rony, Demerjian Jacques and Couturier Raphael, ”On the Performance of Resource-aware Compression Techniques for Vital Signs Data in Wireless Body Sensor Networks” in ”2018 IEEE Middle East & North Africa COMMunications Conference (MENACOMM 2018)”, Jounieh, Lebanon, pp. (6 pages), apr. 2018

4. Tannoury Anthony, Darazi Rony, Makhoul Abdallah and Guyeux Christophe, ”Wireless multimedia sensor network deployment for disparity map calculation” in ”2018

IEEE Middle East & North Africa COMMunications Conference (MENACOMM 2018)", Jounieh, Lebanon, pp. (6 pages), apr. 2018

5. Boudargham Nadine, Bou Abdo Jacques, Demerjian Jacques, Guyeux Christophe and Makhoul Abdallah, "Collaborative body sensor networks: Taxonomy and open challenges" in "2018 IEEE Middle East & North Africa COMMunications Conference (MENACOMM 2018)", pp. (6 pages), apr. 2018

6. Azar Joseph, Darazi Rony, Habib Carol, Makhoul Abdallah and Demerjian Jacques, "Using DWT Lifting Scheme for Lossless Data Compression in Wireless Body Sensor Networks" in "14th International Wireless Communications and Mobile Computing Conference (IWCMC 2018)", St. Raphael, Cyprus, pp. A venir, jun. 2018

7. Moustafa Harb Hassan, Makhoul Abdallah and Abou Jaoude Chady, "En-Route Data Filtering Technique for Maximizing Wireless Sensor Network Lifetime" in "14th International Wireless Communications and Mobile Computing Conference (IWCMC 2018)", St. Raphael, Cyprus, pp. A venir, jun. 2018

8. Koussaifi Maroun, Habib Carol and Makhoul Abdallah, "Real-time Stress Evaluation using Wireless Body Sensor Networks" in "10th IEEE Wireless Days Conference", vol. A venir, apr. 2018

9. Battat Nadia, Makhoul Abdallah, Kheddouci Hamamache, and Medjahed Sabrina, "Trust Based Monitoring Approach for Mobile Ad Hoc Networks" in "16th International Conference on Ad Hoc Networks and Wireless, ADHOC-NOW 2017", Messina, Italy, vol. 10517 de Lecture Notes in Computer Science (LNCS), pp. 55-62, sep. 2017

10. Habib Carol, Makhoul Abdallah, Couturier Raphael and Darazi Rony, "On The Problem of Energy Efficient Mechanisms Based on Data Reduction in Wireless Body Sensor Networks" in "Eleventh International Conference on Sensor Technologies and Applications SENSORCOMM 2017", Rome, Italy, pp. 94 - 98, sep. 2017

11. Tannoury Anthony, Darazi Rony, Guyeux Christophe and Makhoul Abdallah, "Efficient and accurate monitoring of the depth information in a Wireless Multimedia Sensor Network based surveillance" in "1st IEEE International Conference on Sensors, Networks, Smart and Emerging Technologies (SENSET2017)", pp. A venir, sep. 2017

12. Moustafa Harb Hassan, Makhoul Abdallah, Tawbi Samar and Zahwe Oussama, "Energy efficient filtering techniques for data aggregation in sensor networks" in "13th International Wireless Communications and Mobile Computing Conference (IWCMC 2017)", pp. 693 - 698, jun. 2017

13. Habib Carol, Makhoul Abdallah, Darazi Rony and Couturier Raphael, "Real-time Sampling Rate Adaptation based on Continuous Risk Level Evaluation in Wireless Body Sensor Networks" in "13th IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob 2017)", vol. A venir, pp. A venir, oct. 2017

14. Boudargham Nadine, Bou Abdo Jacques, Demerjian Jacques, Guyeux Christophe and Makhoul Abdallah, "Investigating Low Level Protocols for Wireless Body Sensor Networks" in "2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)", Agadir, Morocco , pp. (6 pages), dec. 2016

15. Atoui Ibrahim, Makhoul Abdallah, Tawbi Samar, Couturier Raphael and Hijazi Abbas, ”Tree-based data aggregation approach in periodic sensor networks using correlation matrix and polynomial regression” in ”2016 IEEE Intl Conference on Computational Science and Engineering (CSE)”, pp. 716 - 723, aug. 2016

16. Habib Carol, Makhoul Abdallah, Darazi Rony and Couturier Raphael, ”Multisensor Data Fusion for Patient Risk Level Determination and Decision-support in Wireless Body Sensor Networks” in ”19th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems”, Malta, Malta, pp. 221 - 224, nov. 2016

17. Farhat Ahmad, Makhoul Abdallah, Guyeux Christophe, Tawil Rami, Jaber Ali and Hijazi Abbas, ”On the topology effects in wireless sensor networks based prognostics and health management” in ”19th IEEE International Conference on Computational Science and Engineering (CSE 2016)”, Paris, France, pp. A venir, aug. 2016

18. Salim Christian, Makhoul Abdallah, Darazi Rony and Couturier Raphael, ”Combining Frame Rate Adaptation and Similarity Detection for Video Sensor Nodes in Wireless Multimedia Sensor Networks” in ”IWCMC 2016, 12th Int. Wireless Communications and Mobile Computing Conference”, Paphos, Cyprus, pp. 327 - 332, sep. 2016

19. Atoui Ibrahim, Ahmad Ali, Medlej Maguy, Makhoul Abdallah, Tawbi Samar and Hijazi Abbas, ”Tree-based data aggregation approach in wireless sensor network using fitting functions” in ”ICDIPC 2016, 6th International Conference on Digital Information Processing and Communications”, Beirut, Lebanon, pp. 146 - 150, apr. 2016

20. Salim Christian, Makhoul Abdallah, Darazi Rony and Couturier Raphael, ”Adaptive Sampling Algorithms with Local Emergency Detection for Energy Saving in Wireless Body Sensor Networks” in ”2016 IEEE/IFIP Network Operations and Management Symposium (NOMS)”, Istanbul, Turkey, pp. 745 - 749, apr. 2016

21. Habib Carol, Makhoul Abdallah, Darazi Rony and Couturier Raphael, ”Multisensor Data Fusion and Decision Support in Wireless Body Sensor Networks” in ” 2016 IEEE/IFIP Network Operations and Management Symposium (NOMS)”, Istanbul, Turkey, pp. 708 - 712, apr. 2016

22. Moustafa Harb Hassan, Makhoul Abdallah, Couturier Raphael and Medlej Maguy, ”An Aggregation and Transmission Protocol for Conserving Energy in Periodic Sensor Networks” in ”2015 IEEE 24th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)”, Larnaca, Cyprus, pp. 134 - 139, jun. 2015

23. Al-Sharif Rola, Guyeux Christophe, Fadil Yousra Ahmed, Makhoul Abdallah and Jaber Ali, ”On the usefulness of information hiding techniques for wireless sensor networks security” in ”ADHOCNETS 14, 6th Int. Conf. on Ad Hoc Networks”, Rhodes, Greece, vol. 140, pp. 51-62, aug. 2014

24. Elghers Sabrina, Makhoul Abdallah and Laiymani David, ”Local Emergency Detection Approach for Saving Energy in Wireless Body Sensor Networks” in ”WIMOB

2014, 10th IEEE Int. Conf. on Wireless and Mobile Computing, Networking and Communications", Larnaca, Cyprus, pp. 585-591, oct. 2014

25. Moustafa Harb Hassan, Makhoul Abdallah, Laiymani David, Jaber Ali and Tawil Rami, "K-Means Based Clustering Approach for Data Aggregation in Periodic Sensor Networks" in "WIMOB 2014, 10th IEEE Int. Conf. on Wireless and Mobile Computing, Networking and Communications", Larnaca, Cyprus, pp. 434–441, oct. 2014

26. Moustafa Harb Hassan, Makhoul Abdallah, Tawil Rami and Jaber Ali, "A Suffix-Based Enhanced Technique for Data Aggregation in Periodic Sensor Networks" in "IWCMC 2014, 10th IEEE Int. Wireless Communications and Mobile Computing Conference", Nicosia, Cyprus, pp. 494–499, aug. 2014

27. Guyeux Christophe, Makhoul Abdallah and Bahi Jacques, "A Security Framework for Wireless Sensor Networks: Theory and Practice" in "WETICE 2014, 23rd IEEE WETICE Conference, 4th Track on Cyber Physical Society with SOA, BPM and Sensor Networks", Parma, Italy, pp. 269-274, jun. 2014

28. Laiymani David and Makhoul Abdallah, "Adaptive data collection approach for periodic sensor networks" in "IWCMC 2013, 9th IEEE Int. Wireless Communications and Mobile Computing Conference", Belgrade, Serbia, pp. 1448–1453, jul. 2013

29. Bahi Jacques, Makhoul Abdallah and Medlej Maguy, "Frequency Filtering Approach for Data Aggregation in Periodic Sensor Networks" in "NOMS 2012, 13-th IEEE/IFIP Network Operations and Management Symposium", Hawaii, United States, pp. 570 - 573, apr. 2012

30. Bahi Jacques, Makhoul Abdallah and Medlej Maguy, "An Optimized In-Network Aggregation Scheme for Data Collection in Periodic Sensor Networks" in "ADHOC-NOW 2012, 11-th Int. Conf. on Ad Hoc Networks and Wireless", Belgrade, Serbia, vol. 7363 de Lecture Notes in Computer Science (LNCS), pp. 153–166, jul. 2012

31. Bahi Jacques, Guyeux Christophe, Makhoul Abdallah and Pham Congduc, "Secure scheduling of wireless video sensor nodes for surveillance applications" in "ADHOCNETS 11, 3rd Int. ICST Conference on Ad Hoc Networks", Paris, France, vol. 89 de LNICST, pp. 1–15, sep. 2011

32. Bahi Jacques, Hakem Mourad and Makhoul Abdallah, "Reliable Distributed Data Fusion Scheme in Unsafe Sensor Networks" in "AICCSA 2011, 9-th ACS/IEEE Int. Conf. on Computer Systems and Applications", Sharm El-Sheikh, Egypt, pp. 46-53, dec. 2011

33. Bahi Jacques, Makhoul Abdallah and Medlej Maguy, "Energy Efficient 2-Tiers Weighted in-Sensor Data Cleaning" in "SENSORCOMM'11, 5-th Int. Conf. on Sensor Technologies and Applications", Nice, France, pp. 197–202, aug. 2011

34. Bahi Jacques, Makhoul Abdallah and Medlej Maguy, "Data Aggregation for Periodic Sensor Networks Using Sets Similarity Functions" in "IWCMC 2011, 7th IEEE Int. Wireless Communications and Mobile Computing Conference", Istanbul, Turkey, pp. 559–564, jul. 2011

35. Bahi Jacques, Guyeux Christophe and Makhoul Abdallah, ”Secure Data Aggregation in Wireless Sensor Networks. Homomorphism versus Watermarking Approach” in ”ADHOCNETS 2010, 2nd Int. Conf. on Ad Hoc Networks”, Victoria, Canada, vol. 49 de Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (Lecture Notes in ICST), pp. 344–358, aug. 2010

36. Pham Congduc and Makhoul Abdallah, ”Performance study of multiple cover-set strategies for mission-critical video surveillance with wireless video sensors” in ”WIMOB 2010, 6th IEEE Int. Conf. on Wireless and Mobile Computing, Networking and Communications,”, Niagara Falls, Canada, pp. 208–216, oct. 2010

37. Bahi Jacques, Guyeux Christophe and Makhoul Abdallah, ”Efficient and Robust Secure Aggregation of Encrypted Data in Sensor Networks” in ”SENSORCOMM’10, 4-th Int. Conf. on Sensor Technologies and Applications”, Venice-Mestre, Italy, pp. 472–477, jul. 2010

38. Abdallah Makhoul, Rachid Saadi, CongDuc Pham, ”Coverage and adaptive scheduling algorithms for criticality management on video wireless sensor networks” in ”ICUMT 2009”, pp. 1–8, oct. 2009

39. Abdallah Makhoul, CongDuc Pham, ” Dynamic scheduling of cover-sets in randomly deployed Wireless Video Sensor Networks for surveillance applications” in ”Wireless Days 2009”, pp. 1–6, Paris, oct. 2009

40. Makhoul Abdallah, Bahi Jacques and Mostefaoui Ahmed, ”Improving Lifetime and Coverage Through a Mobile Beacon for High Density Sensor Networks” in ”SENSORCOMM’08, 2nd IEEE Int. Conf. on Sensor Technologies and Applications”, Cap Esterel, France, pp. 335–341, 2008

41. Makhoul Abdallah, Bahi Jacques and Mostefaoui Ahmed, ”Localization and Coverage for High Density Sensor Networks” in ”PerComW’07, 5th IEEE Int. Conf. on Pervasive Computing and Communications Workshops”, New York, United States, pp. 295–300, mar. 2007

42. Makhoul Abdallah, Bahi Jacques and Mostefaoui Ahmed, ”A Mobile Beacon Based Approach for Sensor Network Localization” in ”WiMob’07, 3rd IEEE Int. Conf. on Wireless and Mobile Computing, Networking and Communications”, New York, United States, pp. 44, oct. 2007

BIBLIOGRAPHY

[1] T. Qiu, N. Chen, K. Li, M. Atiquzzaman, and W. Zhao. How can heterogeneous internet of things build our future: A survey. *IEEE Communications Surveys Tutorials*, 2018.

[2] Jennifer Yick, Biswanath Mukherjee, and Dipak Ghosal. Wireless sensor network survey. *Computer Networks*, 52(12):2292 – 2330, 2008.

[3] S. Sharma, R. K. Bansal, and S. Bansal. Issues and challenges in wireless sensor networks. In *2013 International Conference on Machine Intelligence and Research Advancement*, pages 58–62, 2013.

[4] D. Baum and CIO Information Matters. Big data, big opportunity. *http://www.oracle.com/us/c-central/cio-solutions/informationmatters/big-data-big-opportunity/index.html*, 2013.

[5] Siyao Cheng, Zhipeng Cai, Jianzhong Li, and Xiaolin Fang. Drawing dominant dataset from big sensory data in wireless sensor networks. *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 531–539, 2015.

[6] Tongxin Zhu, Siyao Cheng, Zhipeng Cai, and Jianzhong Li. Critical data points retrieving method for big sensory data in wireless sensor networks. *EURASIP Journal on Wireless Communications and Networking*, 2016(1):1–14, 2016.

[7] David Laiymani and Abdallah Makhoul. Adaptive data collection approach for periodic sensor networks. In *Wireless Communications and Mobile Computing Conference (IWCMC), 2013 9th International*, pages 1448–1453. IEEE, 2013.

[8] Abdallah Makhoul, Hassan Harb, and David Laiymani. Residual energy-based adaptive data collection approach for periodic sensor networks. *Ad Hoc Networks*, 35:149–160, 2015.

[9] Abdallah Makhoul and Hassan Harb. Data reduction in sensor networks: Performance evaluation in a real environment. *IEEE Embedded Systems Letters*, 9(4):101–104, 2017.

[10] Gaby Bou Tayeh, Abdallah Makhoul, David Laiymani, and Jacques Demerjian. A distributed real-time data prediction and adaptive sensing approach for wireless sensor networks. *Pervasive and mobile computing*, 49(2018):62–75, 2018.

[11] Jacques Bahi, Abdallah Makhoul, and Maguy Medlej. A two tiers data aggregation scheme for periodic sensor networks. *Ad Hoc & Sensor Wireless Networks*, 21((1-2)):77–100, 2014.

[12] Jacques M. Bahi, Abdallah Makhoul, and Maguy Medlej. An optimized in-network aggregation scheme for data collection in periodic sensor networks. *Ad-hoc, Mobile, and Wireless Networks: 11th International Conference, ADHOC-NOW 2012, Belgrade, Serbia, July 9-11, 2012. Proceedings*, pages 153–166, 2012.

[13] Hassan Harb, Abdallah Makhoul, and Raphaël Couturier. An enhanced k-means and anova-based clustering approach for similarity aggregation in underwater wireless sensor networks. *IEEE Sensors journal*, 15(10):5483–5493, 2015.

[14] Hassan Moustafa Harb, Abdallah Makhoul, David Laiymani, and Ali Jaber. A distance-based data aggregation technique for periodic sensor networks. *ACM Transactions on Sensor Networks*, 13(4):32 (40 pages), sep 2017.

[15] Carol Habib, Abdallah Makhoul, Rony Darazi, and Christian Salim. Self-adaptive data collection and fusion for health monitoring based on body sensor networks. *IEEE Transactions on Industrial Informatics*, 12(6):2342–2352, 2016.

[16] Carol Habib, Abdallah Makhoul, Rony Darazi, and Raphaël Couturier. Multisensor data fusion and decision support in wireless body sensor networks. In *Network Operations and Management Symposium (NOMS), 2016 IEEE/IFIP*, pages 708–712. IEEE, 2016.

[17] Carol Habib, Abdallah Makhoul, Rony Darazi, and Raphaël Couturier. Multisensor data fusion for patient risk level determination and decision-support in wireless body sensor networks. In *Proceedings of the 19th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, pages 221–224. ACM, 2016.

[18] Jacques M. Bahi, Christophe Guyeux, Mourad Hakem, and Abdallah Makhoul. Epidemiological approach for data survivability in unattended wireless sensor networks. *J. Network and Computer Applications*, 46:374–383, 2014.

[19] Abdallah Makhoul, Christophe Guyeux, Mourad Hakem, and Jacques M. Bahi. Using an epidemiological approach to maximize data survival in the internet of things. *ACM Trans. Internet Techn.*, 16(1):5:1–5:15, 2016.

[20] Silvia Santini and Kay Römer. An adaptive strategy for quality-based data reduction in wireless sensor networks. In *Proceedings of the 3rd International Conference on Networked Sensing Systems*, pages 29–36, 2006.

[21] L. Tan and M. Wu. Data reduction in wireless sensor networks: A hierarchical lms prediction approach. *IEEE Sensors Journal*, 16(6):1708–1715, March 2016.

[22] Mou Wu, Liansheng Tan, and Naixue Xiong. Data prediction, compression, and recovery in clustered wireless sensor networks for environmental monitoring applications. *Information Sciences*, 329(Supplement C):800 – 818, 2016.

[23] U. Raza, A. Camerra, A. L. Murphy, T. Palpanas, and G. P. Picco. Practical data prediction for real-world wireless sensor networks. *IEEE Transactions on Knowledge and Data Engineering*, 27(8):2231–2244, Aug 2015.

[24] J. M. C. Silva, K. A. Bispo, P. Carvalho, and S. R. Lima. Litesense: An adaptive sensing scheme for wsns. In *2017 IEEE Symposium on Computers and Communications (ISCC)*, pages 1209–1212, 2017.

[25] Yongjae Jon. Adaptive sampling in wireless sensor networks for air monitoring system. Master's thesis, Uppsala University, Department of Information Technology, 2016.

[26] Jinseok Yang, Sameer Tilak, and Tajana S Rosing. An interactive context-aware power management technique for optimizing sensor network lifetime. In *SENSORNETS*, pages 69–76, 2016.

[27] Yao Liang and Yimei Li. An efficient and robust data compression algorithm in wireless sensor networks. *IEEE Communications Letters*, 18(3):439–442, 2014.

[28] Evangelos Zimos, Dimitris Toumpakaris, Adrian Munteanu, and Nikos Deligiannis. Multiterminal source coding with copula regression for wireless sensor networks gathering diverse data. *IEEE Sensors Journal*, 17(1):139–150, 2017.

[29] Jingfei He, Guiling Sun, Zhouzhou Li, and Ying Zhang. Compressive data gathering with low-rank constraints for wireless sensor networks. *Signal Processing*, 131:73–76, 2017.

[30] H. Wu, J. Wang, M. Suo, and P. Mohapatra. A holistic approach to reconstruct data in ocean sensor network using compression sensing. *IEEE Access*, PP(99):1–1, 2017.

[31] Aseel Basheer and Kewei Sha. Cluster-based quality-aware adaptive data compression for streaming data. *J. Data and Information Quality*, 9(1):2:1–2:33, September 2017.

[32] A. Masoum, N. Meratnia, and P.J.M. Havinga. An energy-efficient adaptive sampling scheme for wireless sensor networks. *8th International Conference on Intelligent Sensors, Sensor Networks and Information Processing, IEEE,*, pages , 231–236, 2013.

[33] A. Wood, G.V. Merrett, S.R. Gunn, B.M. Al-Hashimi, and W. Shadbolt, N.R.and Hall. Adaptive sampling in context-aware systems: a machine learning approach. *IET Conference on Wireless Sensor Systems,*, pages , 1–5, 2012.

[34] B. Gedik, L. Liu, and P. Yu. Asap:an adaptive sampling approach to data collection in sensor networks. *IEEE Transactions on Parallel Distributed Systems,*, 18(12):, 1766–1783, 2007.

[35] M. Vahabi, M.F.A. Rasid, R.S.A.R. Abdullah, and M.H.F. Ghazvini. Adaptive data collection algorithm for wireless sensor networks. *IJCSNS International Journal of Computer Science and Network Security,*, 8(6):, 125–132, 2008.

[36] Ankur Jain, Edward Y. Chang, and Yuan-Fang Wang. Adaptive stream resource management using kalman filters. In *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, SIGMOD ’04, pages 11–22. ACM, 2004.

[37] Silvia Santini and Kay Römer. An adaptive strategy for quality-based data reduction in wireless sensor networks. In *Proceedings of the 3rd International Conference on Networked Sensing Systems (INSS 2006)*, pages 29–36, 2006.

[38] B. Qutub Ali, N. Pissinou, and K. Makki. Approximate replication of data using adaptive filters in wireless sensor networks. In *2008 3rd International Symposium on Wireless Pervasive Computing*, pages 365–369, 2008.

[39] L. Tan and M. Wu. Data reduction in wireless sensor networks: A hierarchical lms prediction approach. *IEEE Sensors Journal*, 16(6):1708–1715, 2016.

[40] Junlei Li, James McCann, Nancy Pollard, and Christos Faloutsos. Dynammo: Mining and summarization of coevolving sequences with missing values. *ACM SIGKDD, June/July 2009, pp 527–534*, (CMU-RI-TR-), June 2009.

[41] Leonardo C. Monteiro, Flavia C. Delicato, Luci Pirmez, Paulo F. Pires, and Claudio Miceli. Dpcas: Data prediction with cubic adaptive sampling for wireless sensor networks. *Green, Pervasive, and Cloud Computing*, pages 353–368, 2017.

[42] V Nagesh Babu and A Arudra. Enhancement of secure and efficient data transmission in cluster based wireless sensor networks. *International Journal of Scientific and Research Publications*, 4(6):1–6, 2014.

[43] Patrick E McKight and Julius Najab. Kruskal-wallis test. *Corsini Encyclopedia of Psychology*, 2010.

[44] Jerry Zhao and Ramesh Govindan. Understanding packet delivery performance in dense wireless sensor networks. In *Proceedings of the 1st International Conference on Embedded Networked Sensor Systems*, SenSys '03, pages 1–13, New York, NY, USA, 2003. ACM.

[45] Yao Liang and Yimei Li. An efficient and robust data compression algorithm in wireless sensor networks. *IEEE Communications Letters*, 18(3):439–442, 2014.

[46] 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*, 24((3-4)):199–217, 2015.

[47] C. Chao, , and T. Hsiao. Design of structure-free and energy-balanced data aggregation in wireless sensor networks. *Journal of Network and Computer Applications*, 17:229–239, 2014.

[48] Ali Norouzi, Faezeh Sadat Babamir, and Z.Orman. A tree based data aggregation scheme for wireless sensor networks using ga. *Wireless Sensor Network*, 4(8):191–196, 2012.

[49] Yao Lu, I.S. Comsa, P. Kuonen, and B. Hirsbrunner. Dynamic data aggregation protocol based on multiple objective tree in wireless sensor networks. *Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), IEEE*, pages 1–7, 2015.

[50] Yung-Kuei Chiang, Neng-Chung Wang, and Chih-Hung Hsieh. A cycle-based data aggregation scheme for grid-based wireless sensor networks. *Sensors*, 14(5):8447–8464, 2014.

[51] Chaonan Wang, Liudong Xing, Vinod M. Vokkarane, and Yan Sun. Reliability of wireless sensor networks with tree topology. *International Journal of Performability Engineering*, 8(2):213–216, 2012.

[52] Ketki Ram Bhakare, R.K. Krishna, and Samiksha Bhakare. An energy-efficient grid based clustering topology for a wireless sensor network. *International Journal of Computer Applications*, 39(14), 2012.

[53] Haydar Abdulameer Marhoon, M. Mahmuddin, and Shahrudin Awang Nor. Chain-based routing protocols in wireless sensor networks: A survey. *ARPN Journal of Engineering and Applied Sciences*, 10(3):1389–1398, 2015.

[54] Pinghui Zou and Yun Liu. A data-aggregation scheme for wsn based on optimal weight allocation. *Journal Of networks*, 9(1):100–107, 2014.

[55] M. Shanmukhi and O.B.V. Ramanaiah. Cluster-based comb-needle model for energy-efficient data aggregation in wireless sensor networks. *Applications and Innovations in Mobile Computing (AIMoC)*, pages 42–47, 2015.

[56] Tao Du, Zhe Qu, Qingbei Guo, and Shouning Qu. A high efficient and real time data aggregation scheme for wsns. *International Journal of Distributed Sensor Networks*, 2015(2015):11 pages, 2015.

[57] Hassan Harb, Abdallah Makhoul, Maguy Medlej, and Raphaël Couturier. An aggregation and transmission protocol for conserving energy in periodic sensor networks. *24th IEEE International Conference Enabling Technologies: Infrastructure for Collaborative Enterprises (Wetice)*, page 2015, 134–139.

[58] Nadeem Javaid, Mohsin Raza Jafri, Zahoor Ali Khan, Nabil Alrajeh, Muhammad Imran, and Athanasios Vasilakos. Chain-based communication in cylindrical underwater wireless sensor networks. *Sensors*, 15:3625–3649, 2015.

[59] C.-M. Chao and T.-Y. Hsiao. Design of structure-free and energy-balanced data aggregation in wireless sensor networks. *Journal of Network and Computer Applications*, 37:229–239, 2014.

[60] Junhai Luo and Jiyang Cai. A dynamic virtual force-based data aggregation algorithm for wireless sensor networks. *International Journal of Distributed Sensor Networks*, 2015(2015):7 pages, 2015.

[61] May Kamil Al-Azzawi, Juan Luo, and Renfa Li. Virtual cluster model in clustered wireless sensor network using cuckoo inspired metaheuristic algorithm. *International Journal of Hybrid Information Technology*, 8(4):133–146, 2015.

[62] Hemavathi Natarajan and Sudha Selvaraj. A fuzzy based predictive cluster head selection scheme for wireless sensor networks. *In Proc. of the 8th International Conference on Sensing Technology*, pages 560–567, 2014.

[63] Dilip Kumar. Performance analysis of energy efficient clustering protocols for maximising lifetime of wireless sensor networks. *IET Wirel. Sensor Syst.*, 4(1):9–16, 2014.

[64] A. Anbarasan, Sivasubramaniam S., and M. Mohanasundhram. A minimum cost effective cluster algorithm using uwsn. *International Journal of Innovative Research in Science, Engineering and Technology*, 3(7):14656–14661, 2014.

[65] K. Ovaliadis and N. Savage. Cluster protocols in underwater sensor networks: a research review. *Journal of Engineering Science and Technology Review*, 7(3):171–175, 2014.

[66] Muhammad Ayaz, Azween Abdullah, Ibrahima Faye, and Yasir Batira. An efficient dynamic addressing based routing protocol for underwater wireless sensor networks. *Computer Communications*, 35(4):475–486, 2012.

[67] Liang Zhao and Qilian Liang. Optimum cluster size for underwater acoustic sensor networks. *Proceeding of the 2006 IEEE conference on Military communications (MILCOM'06)*, pages 1–5, 2006.

[68] Navid Amini, Alireza Vahdatpour, Wenyao Xuand, Mario Gerla, and Majid Sarrafzadeh. Cluster size optimization in sensor networks with decentralized cluster-based protocols. *Computer Communication*, 35(2):207–220, 2012.

[69] Guangsong Yang, Mingbo Xiao, En Cheng, and Jing Zhang. A cluster-head selection scheme for underwater acoustic sensor networks. *Proceeding of International Conference on Communications and Mobile Computing (CMC'10)*, pages 188–191, 2010.

[70] Hyung-Sin Kim, Jin-Seok Han, and Yong-Hwan Lee. Scalable network joining mechanism in wireless sensor networks. *In Proceeding of the IEEE Topical Conference on Wireless Sensors and Sensor Networks (WiSNet'12)*, pages 45–48, 2012.

[71] Abdallah Makhoul, David Laiymani, Hassan Harb, and Jacques M. Bahi. An adaptive scheme for data collection and aggregation in periodic sensor networks. *IJSNet*, 18(1/2):62–74, 2015.

[72] A. Arasu, V. Ganti, and R. Kaushik. Efficient exact set-similarity joins. *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB 2006)*, pages 918–929, 2006.

[73] Hassan Harb, Abdallah Makhoul, Rami Tawil, and Ali Jaber. A suffix-based enhanced technique for data aggregation in periodic sensor networks. *In 10th IEEE Int. Wireless Communications and Mobile Computing Conference (IWCMC)*, pages 494–499, 2014.

[74] Hassan Harb, Abdallah Makhoul, David Laiymani, Ali Jaber, and Rami Tawil. K-means based clustering approach for data aggregation in periodic sensor networks. *In 10th IEEE Int. Conf. on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pages 434–441, 2014.

[75] Hassan Harb, Abdallah Makhoul, David Laiymani, Ali Jaber, and Oussama Bazzi. An analysis of variance-based methods for data aggregation in periodic sensor networks. *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXII (TLDKS)*, 9430:165–183, 2015.

[76] Michel Marie Deza and Elena Deza. Encyclopedia of distances. *Springer (2009)*, pages 1–583, 2009.

[77] Abin Abraham Oommen, C. Senthil Singh, and M. Manikandan. Design of face recognition system using principal component analysis. *International Journal Of Research In Engineering And Technology*, 3(1):6–10, 2014.

[78] Menahem Friedmana, Mark Lastb, Yaniv Makoverb, and Abraham Kandelc. Anomaly detection in web documents using crisp and fuzzy-based cosine clustering methodology. *Information Sciences*, 177:467–475, 2007.

[79] Jun Ye. Cosine similarity measures for intuitionistic fuzzy sets and their applications. *Mathematical and Computer Modelling*, 53((1–2)):91–97, 2011.

[80] Q. Gang, S. Shamik, G. Yuelong, and P. Sakti. Similarity between euclidean and cosine angle distance for nearest neighbor queries. *Proceeding of the 2004 ACM symposium on Applied computing*, pages 1232–1237, 2004.

[81] Samuel Madden. Intel berkeley research lab. *http://db.csail.mit.edu/labdata/labdata.html*, 2004.

[82] Argo. Online data. *http://www.argo.ucsd.edu/index.html*, 2000.

[83] Eduardo F. Nakamura, Antonio A. F. Loureiro, and Alejandro C. Frery. Information fusion for wireless sensor networks: Methods, models, and classifications. *ACM Comput. Surv.*, 39(3), September 2007.

[84] Carmen CY Poon, Benny PL Lo, Mehmet Rasit Yuce, Akram Alomainy, and Yang Hao. Body sensor networks: In the era of big data and beyond. *IEEE reviews in biomedical engineering*, 8:4–16, 2015.

[85] Shahina Begum, Shaibal Barua, and Mobyen Uddin Ahmed. Physiological sensor signals classification for healthcare using sensor data fusion and case-based reasoning. *Sensors*, 14(7):11770–11785, 2014.

[86] Ivan Miguel Pires, Nuno M Garcia, Nuno Pombo, and Francisco Flórez-Revuelta. From data acquisition to data fusion: a comprehensive review and a roadmap for the identification of activities of daily living using mobile devices. *Sensors*, 16(2):184, 2016.

[87] Giancarlo Fortino, Stefano Galzarano, Raffaele Gravina, and Wenfeng Li. A framework for collaborative computing and multi-sensor data fusion in body sensor networks. *Information Fusion*, 22:50–70, 2015.

[88] Shahina Begum, Mobyen Uddin Ahmed, Peter Funk, Ning Xiong, and Bo Von Schéele. A case-based decision support system for individual stress diagnosis using fuzzy similarity matching. *Computational Intelligence*, 25(3):180–195, 2009.

[89] Ya-Li Zheng, Xiao-Rong Ding, Carmen Chung Yan Poon, Benny Ping Lai Lo, Heye Zhang, Xiao-Lin Zhou, Guang-Zhong Yang, Ni Zhao, and Yuan-Ting Zhang. Unobtrusive sensing and wearable devices for health informatics. *IEEE Transactions on Biomedical Engineering*, 61(5):1538–1554, 2014.

[90] Tuba Yilmaz, Robert Foster, and Yang Hao. Detecting vital signs with wearable wireless sensors. *Sensors*, 10(12):10837–10862, 2010.

[91] Giancarlo Fortino, Roberta Giannantonio, Raffaele Gravina, Philip Kuryloski, and Roozbeh Jafari. Enabling effective programming and flexible management of efficient body sensor network applications. *IEEE Transactions on Human-Machine Systems*, 43(1):115–133, 2013.

[92] Nourchène Bradai, Lamia Chaari Fourati, and Lotfi Kamoun. Wban data scheduling and aggregation under wban/wlan healthcare network. *Ad Hoc Networks*, 25:251–262, 2015.

[93] Vijay Raghunathan, Saurabh Ganeriwal, and Mani Srivastava. Emerging techniques for long lived wireless sensor networks. *IEEE Communications Magazine*, 44(4):108–114, 2006.

[94] Hadi Banaee, Mobyen Uddin Ahmed, and Amy Loutfi. Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges. *Sensors*, 13(12):17472–17500, 2013.

[95] Bahador Khaleghi, Alaa Khamis, Fakhreddine O Karray, and Saiedeh N Razavi. Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, 14(1):28–44, 2013.

[96] Bijan Davvaz and Irina Cristea. Fuzzy algebraic hyperstructures: An introduction. Springer 2015.

[97] Health risk assessment and decision-making for patient monitoring and decision-support using wireless body sensor networks. *Information Fusion*, 47:10 – 22, 2019.

[98] Raffaele Gravina, Parastoo Alinia, Hassan Ghasemzadeh, and Giancarlo Fortino. Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges. *Information Fusion*, 35:68–80, 2017.

[99] Saisakul Chernbumroong, Shuang Cang, and Hongnian Yu. A practical multi-sensor activity recognition system for home-based care. *decision support systems*, 66:61–70, 2014.

[100] Saisakul Chernbumroong, Shuang Cang, and Hongnian Yu. Maximum relevancy maximum complementary feature selection for multi-sensor activity recognition. *Expert Systems with Applications*, 42(1):573–583, 2015.

[101] Lei Gao, AK Bourke, and John Nelson. Evaluation of accelerometer based multi-sensor versus single-sensor activity recognition systems. *Medical Engineering and Physics*, 36(6):779–785, 2014.

[102] Muhammad Shoaib, Stephan Bosch, Ozlem Durmaz Incel, Hans Scholten, and Paul JM Havinga. Fusion of smartphone motion sensors for physical activity recognition. *Sensors*, 14(6):10146–10176, 2014.

[103] Boon-Giin Lee and Wan-Young Chung. A smartphone-based driver safety monitoring system using data fusion. *Sensors*, 12(12):17536–17552, 2012.

[104] Hui Wang, Hyeok-soo Choi, Nazim Agoulmine, M Jamal Deen, and James Won-Ki Hong. Information-based energy efficient sensor selection in wireless body area networks. In *Communications (ICC), 2011 IEEE International Conference on*, pages 1–6. IEEE, 2011.

[105] Alexandros Pantelopoulos and Nikolaos G Bourbakis. Prognosis a wearable health-monitoring system for people at risk: Methodology and modeling. *IEEE Transactions on Information Technology in Biomedicine*, 14(3):613–621, 2010.

[106] Daniele Apiletti, Elena Baralis, Giulia Bruno, and Tania Cerquitelli. Real-time analysis of physiological data to support medical applications. *IEEE transactions on information technology in biomedicine*, 13(3):313–321, 2009.

[107] Ragesh GK and K Baskaran. A survey on futuristic health care system: Wbans. *Procedia Engineering*, 30:889–896, 2012.

[108] National early warning score. http://www.rcplondon.ac.uk/resources/national-early-warning-score-news, 2012. Accessed: 2016-5-10.

[109] Christian Salim, Abdallah Makhoul, Rony Darazi, and Raphaël Couturier. Adaptive sampling algorithms with local emergency detection for energy saving in wireless body sensor networks. In *Network Operations and Management Symposium (NOMS), 2016 IEEE/IFIP*, pages 745–749. IEEE, 2016.

[110] Hugh Durrant-Whyte and Thomas C. Henderson. *Springer Handbook of Robotics*, chapter Multisensor Data Fusion, pages 585–610. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[111] Raghavendra V Kulkarni, Anna Förster, and Ganesh Kumar Venayagamoorthy. Computational intelligence in wireless sensor networks: a survey. *Communications Surveys & Tutorials, IEEE*, 13(1):68–96, 2011.

[112] Physionet. https://www.physionet.org/, 2000 - present. Accessed: 2016-10-4.

[113] Real-time analysis of physiological data to support medical applications. Tech. Rep. Available:http://tasmania.polito.it/~daniele/pub/, 2007. Accessed: 2016-10-4.

[114] Michele Magno, Tommaso Polonelli, Filippo Casamassima, Andres Gomez, Elisabetta Farella, and Luca Benini. Energy-efficient context aware power management with asynchronous protocol for body sensor network. *Mobile Networks and Applications*, pages 1–11, 2016.

[115] Maroun Koussaifi, Carol Habib, and Abdallah Makhoul. Real-time stress evaluation using wireless body sensor networks. *10th IEEE Wireless Days Conference*, apr 2018.

[116] R.M. Anderson and R.M. May. Population biology of infectious disease. *I Nature*, 180:361–367, 1999.

[117] L.-X. Yang, X. Yang, J. Liu, Q. Zhu, and C. Gan. Epidemics of computer viruses: A complex-network approach. *Appl Math Comput*, 219(16):8705–8717, 2013.

[118] Q. Zhu, X. Yang, L.-X. Yang, and J. Ren. Modeling and analysis of the spread of computer virus. *Commun Nonlinear Sci Numer Simu*, 17(2012):5117–5124, 2012.

[119] Q. Zhu, X. Yang, L.-X. Yang, and C. Zhang. Optimal control of computer virus under a delayed model. *Appl Math Comput*, 218(23):11613–11619, 2012.

[120] C. Zhang, Y. Zhao, and Y. Wu. An impulse model for computer viruses. *Discrete Dyn Nat Soc*, 2012, 2012.

[121] C. Zhang, Y. Zhao, Y. Wu, and S. Deng. A stochastic dynamic model of computer viruses. *Discrete Dyn Nat Soc*, 2012(2012).

[122] X. Yang and L.-X. Yang. Towards the epidemiological modeling of computer viruses. *Discrete Dyn Nat Soc*, 2012, 2012.

[123] L.-X. Yang and X. Yang. The spread of computer viruses under the influence of removable storage devices. *Appl Math Comput*, 219(8):3914–3922, 2012.

[124] L.-X. Yang, X. Yang, L. Wen, and J. Liu. A novel computer virus propagation model and its dynamics. *Int J Comput Math*, 89(17):2307–2314, 2012.

[125] L.-X. Yang, X. Yang, Q. Zhu, and L. Wen. A computer virus model with graded cure rates. *Nonlinear Anal: Real World Appl*, 14(1):414–422, 2013.

[126] M. Yang, Z. Zhang, Q. Li, and G. Zhang. An slbrs model with vertical transmission of computer virus over the interne. *Discrete Dyn Nat Soc*, 2012, 2012.

[127] Chenquan Gan, Xiaofan Yang, Wanping Liu, Qingyi Zhu, Jian Jin, and Li He. Propagation of computer virus both across the internet and external computers: A complex-network approach. *Communications in Nonlinear Science and Numerical Simulation*, 19(8):2785–2792, 2014.

[128] W. O. Kermack and Ag McKendrick. A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772):700–721, August 1927.

[129] W.O. Kermack and A.G. McKendrick. Contributions of mathematical theory to epidemics. *Proc. Royal Soc. London – Series A 141*, pages 94–122, 1933.

[130] C.C. Zou, W.B. Gong, D. Towsley, and L.X. Gao. The monitoring and early detection of internet malicious codes. *IEEE/ACM Trans. Network*, 13(5):961–974, 2005.

[131] M.J. Keeling and K.T.D. Eames. Network and epidemic models. *J. R. Soc. Interface*, 2(4):295–307, 2005.

[132] Erol Gelenbe, Varol Kaptan, and Yu Wang. Biological metaphors for agent behavior. *19th International Symposium Computer and Information Sciences – ISCIS 2004*, 3280:667–675, 2004.

[133] M.E.J. Newman, S. Forrest, and J.Balthrop. Email networks and the spread of computer virus. *Phys. Rev. E*, 66:232–369, 2002.

[134] B.K. Mishra and S.K. Pandey. Dynamic model of worms with vertical transmission in computer network. *Appl. Math. Comput*, 217(21):8438–8446, 2011.

[135] M. Draief, A. Ganesh, and L. Massouili. Thresholds for virus spread on network. *Ann. Appl. Probab*, 18(2):359–369, 2008.

[136] B.K. Mishra and N. Jha. Seiqrs model for the transmission of malicious objects in computer network. *Appl. Math. Model*, 34:710–715, 2010.

[137] T. Chen and N. Jamil. Effectiveness of quarantine in malicious codes epidemic. *IEEE International Conference on Communications (ICC)*, page 2142–2147, 2006.

[138] M.E. Alexander, S.M. Moghadas, P. Rohani, and A.R. Summers. Modeling the effect of a booster vaccination on disease epidemiology. *J. Math. Biol*, 52:290–306, 2006.

[139] W.T. Richard and J.C. Mark. Modeling virus propagation in peer-to-peer networks. *IEEE International Conference on Information, Communication and Signal Processing*, pages 981–985, 2005.

[140] Bimal Kumar Mishra and Samir Kumar Pandey. Dynamic model of worm propagation in computer network. *Applied Mathematical Modelling*, 38(7-8):2173–2179, 2014.

[141] Z. Wang, D.W.C. Ho, H. Dong, and H. Gao. Robust h-infinity finite-horizon control for a class of stochastic nonlinear time-varying systems subject to sensor and actuator saturations. *IEEE Trans. Automat. Control*, 55(7):1716–1722, 2010.

[142] B. Shen, Z. Wang, and Y.S. Hung. Distributed consensus h-infinity filtering in sensor networks with multiple missing measurements: the finite-horizon case. *Automatica*, 55(7):1682–1688, 2010.

[143] Roberto Di Pietro and Nino Vincenzo Verde. Epidemic data survivability in unattended wireless sensor networks. pages 11–22, 2011.

[144] Roberto Di Pietro and Nino Vincenzo Verde. Epidemic theory and data survivability in unattended wireless sensor networks: Models and gaps. *Pervasive and Mobile Computing*, 9(4):588 – 597, 2013.

[145] Roberto Di Pietro, Luigi V. Mancini, Claudio Soriente, Angelo Spognardi, and Gene Tsudik. Catch me (if you can): Data survival in unattended sensor networks. pages 185–194, 2008.

[146] Herbert W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42:599–653, 2000.

Abstract:

Wireless sensor networks (WSN) produce a huge amount of data that needs to be gathered, processed, and transmitted according to the application objectives. This document describes data collection and processing in large scale WSN from data acquisition to data fusion and survivability. Our main objectives are to reduce the amount of data traffic, filter redundant measurements, and make predictions and inferences in a WSN with energy, storage capacity, computing power and communications bandwidth resource constraints. Data collection and prediction algorithms are proposed which significantly reduce the size of the collected and transmitted data by adapting the sampling and transmission rates of the sensors. Then, we present three data aggregation techniques to eliminate redundant data sets generated by neighbouring nodes. In addition, we provide information fusion techniques by exploiting the synergy among the collected data. They are based on fuzzy logic and aim to obtain information of greater quality and make accurate decisions. An epidemic model based approach for data survivability is then studied. It consists on preserving data for a long period of time in unattended WSN. Finally, the proposed methods were evaluated on real-world data sets collected at our laboratory and compared to recent related approaches. The results were promising in quality of data collection and transmission reduction.

Keywords: WSN, Distributed algorithms, Adaptive sampling, Data prediction, K-means clustering, Sets Similarity and Distance functions, Information fusion, Fuzzy logic, Data survival, Epidemic models.

Résumé :

Les réseaux de capteurs produisent une très grande quantité de données qui doivent être collectées, traitées et transmises selon les besoins de l'application. Ce document décrit la collecte et le traitement de données dans les réseaux de capteurs à grande échelle. L'objectif principal est de réduire la quantité de données collectées et transmises, de filtrer les mesures redondantes et de faire des prédictions et inférences dans un réseau ayant faibles ressources énergétiques et de stockage. Des algorithmes de collecte de données et de prédiction sont proposés. Ils réduisent considérablement la taille de données collectées en adaptant les taux d'échantillonnage des capteurs. Ensuite, nous présentons trois techniques d'agrégation de données pour éliminer les ensembles de données redondants générés par les nœuds voisins. Par ailleurs, nous proposons des techniques de fusion de données en exploitant la synergie entre les données collectées. Elles sont basées sur la logique floue et visent à obtenir des informations de meilleure qualité et à prendre des décisions plus précises. Une approche de survie de données basée sur les modèles épidémiques est ensuite étudiée. Elle consiste à conserver les données pendant une longue période et durant l'absence du puits. Enfin, les méthodes proposées ont été testées sur des ensembles de données réelles collectées dans notre laboratoire et comparées à des approches existantes. Les résultats étaient prometteurs en termes de la réduction de la taille de données collectées et transmises.

Mots-clés : Réseaux de capteurs, algorithmes distribués, collecte et prédiction de données, analyse de la vairance, agrégation et fusion de données, fonctions de similarité, partitionnement en k-moyennes, système d'inférence floue, survie de données, modèles épidémiologiques.

■ École doctorale SPIM 16 route de Gray F - 25030 Besançon cedex
■ tél. +33 (0)3 81 66 66 02 ■ **ed-spim@univ-fcomte.fr** ■ www.ed-spim.univ-fcomte.fr