THÈSE

En vue de l'obtention du

DIPLOME De DOCTORAT

Délivré par :

L'Université de Limoges, France **&** **L'Université de Tunis El Manar, Ecole Nationale d'Ingénieurs de Tunis, Tunisie**

Disciplines :

Sciences et technologies de l'information et de la communication **Systèmes de Communications**

Présentée et soutenue par :

Manel KORTAS

Optimisation de la liaison montante pour un réseau de capteurs sans fil avec la contrainte d'énergie

Soutenue le 18 juillet 2020 devant les Jury :

Président :	M. Taoufik AGUILI	Professeur à l'ENIT-Tunis, Tunisie
Rapporteurs :	M. Ridha BOUALLEGUE	Professeur au Sup'Com-Tunis, Tunisie
	M. Samir SAOUDI	Professeur à l'IMT Atlantique, France
Examinateurs :	M. Rabah ATTIA	Professeur à l'EPT-Tunis, Tunisie
	M. Oussama HABACHI	Maître de conférences à l'Université de Limoges, France
Directeurs :	M. Vahid MEGHDADI	Professeur à l'Université de Limoges, France
	M. Tahar EZZEDINE	Professeur à l'ENIT-Tunis, Tunisie
	M. Ammar BOUALLEGUE	Professeur émérite à l'ENIT-Tunis, Tunisie

ED 610 - Sciences et Ingénierie des Systèmes, Mathématiques, Informatique

Institut de Recherche : Xlim Systèmes & Réseaux intelligents

Ecole Doctorale : Sciences et Techniques de l'Ingénieur

Laboratoire de Recherche : Systèmes de Communications

THESIS

With a view to obtaining the

DOCTORAL DEGREE

Delivered by:

University of Limoges, France **&** **University of Tunis El Manar, National Engineering school of Tunis, Tunisia**

Disciplines:

Information and Communication Science and technology communications systems **Communications' systems**

Presented by:

Manel KORTAS

Energy Optimization of the uplink for a wireless sensor network with the energy constraint

Presented on July 18, 2020 in front of the Jury:

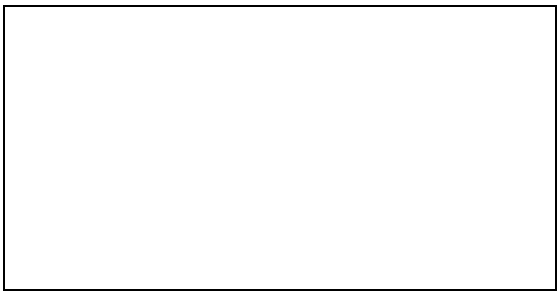
President :	Mr.Taoufik AGUILI	Professor at ENIT-Tunis, Tunisia
Reporters :	Mr.Ridha BOUALLEGUE	Professor at Sup'Com-Tunis, Tunisia
	Mr.Samir SAOUDI	Professor at IMTAtlantique, France
Examiners :	Mr. Rabah ATTIA	Professor at EPT-Tunis, Tunisia
	Mr. Oussama HABACHI	University Lecturer at University of Limoges, France
Directors :	Mr.Vahid MEGHDADI	Professor at University of Limoges, France
	Mr. Tahar EZZEDINE	Professor at ENIT-Tunis, Tunisia
	Mr. Ammar BOUALLEGUE	Professor Emeritus at ENIT-Tunis, Tunisia

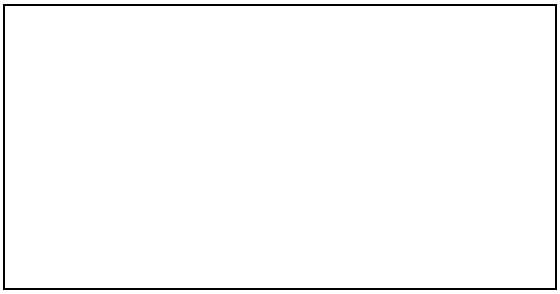
ED 610 – Systems Science and engineering, Mathematics, Computer Science

Research Institute: Xlim Systems &Smart Networks

Doctoral school: Engineer's Science and Techniques

Research Laboratory: Communications Systems





To my source of inspiration and success,
To the best family in the world.
May they find here the testimony of an eternal love.

Abstract

In this dissertation, we are interested in the data gathering with energy constraint for Wireless Sensor Networks (WSNs). Yet, there exist several challenges that may disturb a convenient functioning of this kind of networks. Indeed, WSNs' applications have to deal with limited energy, memory and processing capabilities of sensor nodes. Furthermore, as the size of these networks is growing continually, the amount of data for processing and transmitting becomes enormous. In many practical cases, the wireless sensors are distributed across a physical field to monitor physical phenomena with high space-time correlation. Hence, the main focus of this thesis is to reduce the amount of processed and transmitted data in the data gathering scenario.

In the first part of this thesis, we consider the Compressive Sensing (CS), which is a promising technique to exploit this correlation in order to limit the number of transmission and therefore increase the lifetime of the network. Typically, we are interested in the mesh network topology, where the sink node is not in the range of sensors and routing schemes must be applied. We propose a joint Space-Time Compressive Sensing (STCS) by exploiting jointly the inter-sensors and intra-sensor data dependency. Moreover, since the routing and the number of retransmission affect significantly the total energy consumption, we introduce the routing in our cost function in order to optimize the selection of the transmitting sensors. Simulation results show that this method outperforms the existing ones and confirm the validity of our approach.

In the second part of this thesis, we attempt to address nearly the same twofold energy saving scheme that is investigated in the first part with the use of the Matrix Completion (MC) methodology. Precisely, we assume that a restricted number of sensor nodes are selected to be active and represent the whole network, while the rest of nodes remain idle and do not participate at all in the data sensing and transmission. Furthermore, the set of active nodes' readings is efficiently reduced, in each time slot, according to a cluster scheduling with the Optimized Cluster-based MC data gathering approach (OCBMC). Relying on the existing MC techniques, the sink node is unable to recover the entire data matrix due to the existence of the completely empty rows that correspond to the inactive nodes. Thereby, we propose a complementary

interpolation technique, based on a minimization problem, that benefits from nodes inter-correlation, to guarantee the reconstruction of all the empty rows, despite their large number. The proposed three-stage MC-based reconstruction pattern, combined with the aforementioned data sampling one, is evaluated under extensive simulations. The results confirm the validity of each building block as well as the efficiency of the whole unified structured approach and prove that it outperforms the baseline schema.

Generally, in the WSNs, ensuring long-term survival of the wireless sensor devices is crucial, especially for the non energy harvesting networks. Thus, there is a huge need to further optimize the use of WSN resources. Although applying a high data compression ratio extremely reduces the overall network energy consumption, the network lifetime is not necessarily extended due to the uneven energy depletion of the sensor nodes' batteries. To this end, in the third part of this thesis, we have developed the Energy-Aware Matrix Completion based data gathering approach (EAMC), which designates the active nodes according to their residual energy levels. Furthermore, since we are mainly interested in the high data loss scenarios, the limited amount of delivered data must be sufficient in terms of informative quality it holds in order to reach a good and satisfactory recovery accuracy for the entire network data. For that reason, the EAMC selects the nodes that can best represent the network depending on their inter-correlation as well as the network energy efficiency, with the use of a combined energy-aware and correlation-based metric. This introduced active node cost function changes with the type of application one wants to perform, with the intention to reach a longer lifespan for the network. Therewith, relying on the three-stage MC based approach for data recovery, the proposed scheme achieves an attractive and competitive trade-off between the data reconstruction quality and the network lifetime for all the investigated scenarios.

Résumé

Dans cette thèse, nous nous intéressons à la collecte de données avec la contrainte d'énergie pour les réseaux de capteurs sans fil (RCSFs). En effet, il existe plusieurs défis qui peuvent perturber le bon fonctionnement de ce type de réseaux. Par exemple, les applications des RCSFs doivent faire face aux capacités très limitées en termes d'énergie, de mémoire et de traitement des nœuds de capteurs. De plus, à mesure que la taille de ces réseaux continue de croître, la quantité de données à traiter et à transmettre devient énorme. Dans de nombreux cas pratiques, les capteurs sans fil sont répartis sur un champ physique afin de surveiller les phénomènes physiques à forte corrélation spatio-temporelle. Par conséquent, l'objectif principal de cette thèse est de réduire la quantité de données traitées et transmises dans le scénario de collecte de données.

Dans la première partie de cette thèse, nous utilisons le *Compressive Sensing (CS)*, une technique prometteuse pour exploiter cette corrélation afin de limiter le nombre de transmissions et ainsi augmenter la durée de vie du réseau. En règle générale, nous nous intéressons à la topologie de réseau maillé, où le point de collecte de données n'est pas situé dans le rayon de communication du capteur transmetteur et des schémas de routage doivent être alors appliqués. Nous proposons le *Space-Time Compressive Sensing (STCS)* en exploitant conjointement la dépendance de données inter-capteurs et intra-capteur. De plus, comme le routage et le nombre de retransmissions affectent de manière significative la consommation totale d'énergie, nous introduisons le routage dans notre fonction de coût afin d'optimiser la sélection des capteurs de transmission. Les simulations montrent que cette méthode surpasse les méthodes existantes et confirment la validité de notre approche.

Dans la deuxième partie de cette thèse, nous tentons de traiter un désign d'économie d'énergie presque similaire à celui proposé dans la première partie avec l'utilisation de la méthodologie de *Matrix Completion (MC)*. Précisément, nous supposons qu'un nombre limité de nœuds de capteurs sont sélectionnés pour être actifs et représenter l'ensemble du réseau, tandis que les autres nœuds restent inactifs et ne participent pas du tout à la détection et à la transmission de leurs données. En outre, l'ensemble

de lectures de données des nœuds actifs est efficacement réduit, à chaque intervalle de temps, conformément à une planification de *cluster* avec l'approche de collecte de données *Optimized Cluster-based MC (OCBMC)*. En se basant sur les techniques existantes de *MC*, le point de collecte de données n'est pas en mesure de récupérer l'intégralité de la matrice de données en raison de l'existence de lignes complètement vides correspondant aux nœuds inactifs. Ainsi, nous proposons une technique d'interpolation complémentaire, basée sur un problème de minimisation, qui bénéficie de l'inter-corrélation entre les noeuds de capteurs, afin de garantir la reconstruction de toutes les lignes vides, malgré leur grand nombre. Le modèle *three-stage MC-based reconstruction* proposé, combiné à celui de l'échantillonnage/compression des données susmentionné, est évalué avec des simulations approfondies. Les résultats confirment la validité de chaque bloc constitutif ainsi que l'efficacité de toute l'approche structurée et unifiée et prouvent qu'elle surpasse le schéma le plus proche.

Généralement, dans les RCSFs, il est crucial d'assurer la survie à long terme des capteurs sans fil, en particulier pour les réseaux sans récupération d'énergie. Ainsi, il existe un énorme besoin d'optimiser davantage l'utilisation des ressources énergétique du réseau. Bien que l'application d'un taux de compression des données élevé réduit considérablement la consommation d'énergie globale du réseau, la durée de vie du réseau n'est pas nécessairement prolongée en raison de l'épuisement inégal des batteries des nœuds de capteurs. A cette fin, dans la troisième partie de cette thèse, nous développons l'approche de collecte de données *Energy-Aware Matrix Completion (EAMC)*, qui désigne les nœuds actifs en fonction de leurs niveaux d'énergies résiduelles. De plus, étant donné que nous sommes principalement intéressés par les scénarios de perte de données élevées, la quantité limitée de données fournies doit être suffisante en termes de qualité informative qu'elle détient afin d'atteindre une précision de récupération bonne et satisfaisante pour l'ensemble des données du réseau. Pour cette raison, l'*EAMC* sélectionne les nœuds qui peuvent représenter le mieux le réseau en fonction de leur inter-corrélation ainsi que de l'efficacité énergétique du réseau, avec l'utilisation d'une métrique combinée qui est éco-énergétique et basée sur la corrélation. Cette fonction de coût, qu'on a introduit, change avec le type d'application que l'on veut effectuer, dans le but d'atteindre une durée de vie plus longue pour le réseau. Sur ce, en s'appuyant sur l'approche *three-stage MC-based reconstruction* pour la récupération des données, le schéma proposé permet un compromis attractif

et compétitif entre la qualité de la reconstruction des données et la durée de vie du réseau pour tous les scénarios étudiés.

Acknowledgments

Thanks to Allah, to whom all praise goes, this work has been accomplished. He has guided me and provided me the strive and the power to undertake this dissertation and reach success.

I am grateful to my principal supervisors Mr. Vahid Meghdadi and Mr. Tahar Ezzedine for giving me the chance to do this PhD and affording me the opportunities and the good conditions to succeed. I also thank them for seeing a potential in me and for their continual encouragement and support.

I am also grateful to my supervisor Mr. Ammar Bouallegue for his wide-ranging theoretical expertise that aided me to expand my research domain. I would like also to think him for encouraging me notably throughout times of difficulties, I used to go to his office looking for insightful comments and positivism.

I wish to express my sincere gratitude to Mr. Oussama Habachi for strengthening me and helping me to pursue this thesis and achieve a solid research path towards this work. Special thanks go to his consistent suggestions, constructive comments and precious guidance afforded to improve the quality of this work. To him, I am eternally grateful for the professional and personal supports.

I would express my warmest thanks to the members of the jury. I thank Mr. Taoufik Aguili, professor at National School of Engineers of Tunis, Tunisia, for accepting to chair my jury. I would like to thank Mr. Ridha Bouallegue, professor at Higher School of Communications of Tunis, University of Carthage, Tunisia, and Mr. Samir Saoudi, professor at IMT-Atlantique, France, for accepting to evaluate my thesis work. I also thank Mr. Rabah Attia, professor at Polytechnic School of Tunis, Tunisia, for accepting being my thesis examiner.

I wish to express my deepest gratitude to all my family members; my parents and grand-parents for their immeasurable support during my life and education, my

husband for his endless belief in me and his far-seeing guidance, my sister and brother for their supports during hard times and my beloved son for bringing joy to my life.

Finally, my sincere thanks are addressed to my friends who have helped and supported me to a greater or lesser extent throughout my work.

Contents

List of Figures

Acronyms

5G	**5**th **G**eneration mobile communication system
BP	**B**asis **P**ursuit
BS	**B**ase **S**tation
CB-CS	**C**ovariogram **B**ased **C**ompressive **S**ensing
CBMC	**C**luster-**B**ased **M**atrix **C**ompletion data gathering approach
CDF	**C**umulative **D**istribution **F**unction
CS	**C**ompressive **S**ensing
CSI	**C**hannel **S**tate **I**nformation
DCT	**D**iscrete **C**osine **T**ransform
DFT	**D**iscrete **F**ourier **T**ransform
EAMC	**E**nergy-**A**ware **M**atrix **C**ompltion based data gathering approach
ECB-DNS	**E**nhanced **C**orrelation **B**ased **D**eterministic **N**ode **S**election
EHWSNs	**E**nergy-**H**arvesting **W**ireless **S**ensor **N**etworks
HTC	**H**uman-**T**ype **C**ommunications
IoT	**I**nternet **o**f **T**hings
IS	**I**solated **S**ensor nodes
KLT	**K**arhunen-**L**oève **T**ransform
LMaFit	**L**ow rank **Ma**trix **Fit**ting
MC	**M**atrix **C**ompletion
MIMO	**M**ulti-**I**nput **M**ulti-**O**utput
MIP	**M**utual **I**ncoherence **P**roperty
MSE	**M**ean **S**quared **E**rror
MTC	**M**achine **T**ype **C**ommunications
NUS	**N**onuniform **S**ampler
OCBMC	**O**ptimized **C**luster-**B**ased **M**atrix **C**ompletion data gathering approach
OMP	**O**rthogonal **M**atching **P**ursuit
PCA	**P**rincipal **C**omponent **A**nalysis
RIP	**R**estricted **I**sometry **P**roperty
SDP	semi**d**eFinite **p**rogramming
SRMF	**S**parsity **R**egularized **M**atrix **F**actorization
SRSVD	**S**parsity **R**egularized **SVD**
STCS	**S**pace-**T**ime **C**ompressive **S**ensing
STCS-RA	**S**pace-**T**ime **C**ompressive **S**ensing **R**outing-**A**ware
SVD	**S**ingular **V**alue **D**ecomposition
SVT	**S**ingular **V**alue **T**hresholding
TDD	**T**ime **D**ivision **D**uplexing
WSN	**W**ireless **S**ensor **N**etwork

Part I

Introduction and related works

Chapter 1

Introduction

Contents

1.1 Research Context

With the rapid progress achieved in the information technology fields, the Internet of Things (IoT) has emerged as a new business model composed of billions of connected devices. Hence, it has gained much attention in both industry and scientific community, since it promises to revolutionize the life quality society and industries by bridging the gap between the physical and the digital word. According to [1], it is estimated that, by 2025, around 41.6 billion sensor-based devices, generating 79.4 ZB of Data, will be connected to the Internet as part of the IoT. On the other hand, we expect that the global IoT market will grow from US $190 billion in 2018 to US $1.1 trillion in 2026 [2]. The current need for Machine-Type Communications (MTC) has led to a variety of communication technologies in order to satisfy the heterogeneous IoT requirements[1] [4]. Recently, the massive IoT access has been considered as a part of the 5^{th} generation mobile communication system (5G). Nevertheless, researchers, scientists, and engineers are facing emerging challenges to effectively incorporate the IoT based systems, especially the resource allocation, in the 5G [5]. In fact, the inclusion of the IoT into the 5G and their evolution still represent a formidable technical challenge due to the huge number of sensors and the generated information. Note that one of the main challenges of the 5G is the massive connectivity for MTC and the management of its coexistence with the high data-rate continuous traffic generated by Human-Type Communications (HTC) in an efficient and effective manner.

Wireless Sensor Networks (WSNs), which represent a key pillar of IoT, take place in the center of this revolution. Typically, these networks consist of a large set of sensor nodes that are self-organising and geographically distributed across the monitored area. Despite the miniaturization of these sensor-based devices, they are able to probe different magnitudes. Indeed, they are usually deployed to supervise various physical phenomena with a high resolution and at a low cost, such as in forests, under water and in civilian and habitat application areas [6]. In usual data gathering techniques, each sensor node takes measures and sends periodically its raw data to the sink, which is the collector node, via multi-hop transmission. If nodes face packet losses, due to collisions or buffer overflows, packets are retransmitted, which leads to

[1] Specifically, it aims to automate as much as possible the data communications between devices, in such a way that these latters can occur rightly without any human intervention [3].

a high cost and a heavy traffic. Nevertheless, this kind of data collecting is either impossible or impractical, especially for the large-scale networks, due to the energy and memory limitations of nodes. In fact, these tiny devices operate in an unattended mode and are usually unable to renew their batteries. Hence, reducing the network energy consumption while gathering and forwarding sensory data is the main challenge for these networks since it directly affects their lifetimes and thus their sustainability. This can be achieved by minimizing the amount of information to be communicated. Indeed, establishing energy-efficient data gathering and acquisition schemes, while obviously keeping a good quality in the recovered data, is always welcomed.
The spatial and the temporal correlations that characterize most of the WSNs signal profiles represent a key for the adaptive and efficient data gathering schemes. While the temporal correlation, reflecting the intra-sensor dependency, finds out the time evolution of the signal, the spatial correlation, reflecting the inter-sensors dependency, captures the spatial variation of the signal between the different sensed locations of the network. Benefiting from this property, sensors' resources can be further saved by eliminating the useless and redundant information.

1.2 Problems Statement

In this thesis, we investigate the challenging scenario of data gathering in WSNs. Most of related works have considered the resource access problems like data collisions, losses and re-transmissions. Recently, some researches have focused on the scheduling of data collecting strategies, through the use of compressive sensing. Moreover, the burgeoning demand of many recent applications to deploy more sensor nodes, with their crucial nature of limited power and computational capacities, urges for the establishment of energy-efficient data gathering and acquisition schemes in order to save as longer as possible the sensors' limited batteries.
Usually, the activities for which a sensor node consumes its energy are sensing, processing, and data communication. Most of the existing energy management strategies assume that radio transmission and reception, and in some setups the acquisition/sampling, are the most energy-consuming operations [7]-[9]. Several papers have mainly focused on data compression in order to minimize the energy consumption by reducing the packets size, such as transform coding or entropy coding [10]-[12]. However, these

kinds of in-network processing-based compression schemes require full data signal, and afterwards most of the information is thrown away at the compression stage. Furthermore, they require explicit computational and communication overheads leading to a high space and time complexity at the sensor side, which is preferable to avoid in this type of networks.

In parallel with the consideration of sensors' resources, a second factor, of prime importance, to be taken into account, is the quality of the decompressed data and the accuracy of the missing data recovery in data loss scenarios. Indeed, after receiving the compressed data, the sink node should perform data reconstruction algorithms. That being the case, the purpose of this thesis is to reduce drastically the amount of data readings, while ensuring a sufficiently good data recovery quality at the sink node.

In addition to the minimization of the sensors energy consumption, preserving an energy load balancing between nodes in order prolong the overall network lifetime is another big challenge to tackle. Indeed, due to the multi-hop systems configuration that most of the WSNs adopt, energy consumption between nodes is uneven, leading usually to fast batteries depletion of some nodes, typically the ones that are situated around the sink. Generally, the supervised environments are of harsh nature, which makes the re-change of the exhausted batteries either impractical or a costly task. Therefrom, to ensure a long term monitoring and enhance the network lifespan, there is a need to provide a suitable energy-aware based data gathering technique. Note that the case of rechargeable power supplies of the Energy-Harvesting Wireless Sensor Networks (EHWSNs), where sensor devices can replenish their batteries with energy from the surrounding environment, is out of scope of this thesis, but can be an underlying technique to prolong the lifespan of WSNs.

Motivations

As it is well-known, the principle key that underlies the data sampling techniques and the analog-to-digital conversion in the current used consumer devices is the Nyquist-Shannon sampling theorem [13]. This theorem reports that if the signal sampling rate represents at least twice its maximum frequency component (i.e the so-called Nyquist rate), the recovery process of that signal can be acheived successfully. However, usually in the resource-limited sensors, the signal samples acquisition is specifically followed by

a data compression phase, where the gathered information has to be encoded in a reduced size manner. Accordingly, a substantial portion of the expensively acquired data is eventually thrown away at the compression stage prior to storage or transmission. Moreover, in several emerging applications that we can face in the WSNs, the Nyquist rate is still very high regarding the network capability [14]. Hopefully, under certain conditions, a new paradigm called the Compressive Sensing (CS), or the Compressed Sampling, goes against the common and known wisdom in data acquisition [13], and states that a perfect reconstruction of the whole data may be possible using a number of measurments or data samples that are far lower than those required by the traditional methods (i.e. the rate that respects the Nyquist property) [15]. Particularly, analog CS, denoted also by the Low-rate CS, violates the conventional sampling notion and allows to sample the signal nonuniformly and at a sub-Nyquist frequency [16] [17], permitting to realize savings on the number of data samples to be gathered. Roughly speaking, instead of sampling the compressible signals at the Nyquist frequency and then performing a compression algorithm, the aforementioned presented technique captures them directly in a compressed form using a sub-Nyquist frequency[2], i.e. a simultaneous sampling and compression mecanism. Here, the resulting data measurements don't need to be manipulated for processing in any way before being delivered, except some quantization eventually [18]. To this end, for the case of WSN applications, the principal asset of the CS technique is its common and simple encoding phase [16]. Note that a data vector might hold many small elements and few large ones, in such a way that most of the data signal information is carried by the larger coefficients. Such a data vector is known to be a compressible signal [19]. Among the conditions that one must afford to ensure a "perfect" reconstruction after the CS, we have the signal sparsity feature. That is, the data vector to be processed should hold only a few non-zero elements[3]. Since correlation structure and redundancy that characterize most of the WSNs' signal profiles are often synonymous with sparsity, the CS method seems to be a good fit for such data gathering frameworks. Afterwards, a data reconstruction algorithm is executed at the sink node, who has less energy and computational constraints. Hence, the computation complexity is moved from sensor nodes to the sink. This meets well the resource-constrained devices of WSNs and sig-

[2]In this dissertation, the "sub-Nyquist data acquisition feature" performed with CS refers to measuring and sensing an analog source, by reducing the measurements' projections. The latters denote the discrete-time data measurements obtained from (2.2).

[3]More details are available in 2.1.1.

nificantly reduces their energy consumption. Indeed, unlike the measurement phase, the recovery phase of CS requires a lot of calculation. For that reason, we assume that the collector node possesses the necessary resources to execute the data recovery operation as it is far less constrained compared to the low powered sensor nodes [20].

1.3 Key Contributions

In this thesis, we focus on the data gathering task in the WSNs, and we seek for a good trade-off between the network limited-resources constraint and the end-user requirements. In fact, we have investigated the following questions; ***how to efficiently reduce the number of data readings to be gathered by sensor nodes, while being able to recover the missing ones?*** and ***how to accomplish this task with a near-optimal utilization of sensors resources in order to further extend the overall network lifetime ?***
Trying to provide answers to these questions, in this work, we have developed adaptive and energy-efficient distributed data gathering schemes, each of which is accompanied by the suitable data reconstruction framework. The preponderant work of this thesis is assembled in the following three contributions;

In our first contribution, we address the first question. We take advantage of the spatial and temporal correlations to perform simultaneously both the distributed CS and the local CS, where a subset of well designated active sensors are deterministically chosen to be representative of the network for the entire detection period, and to gather measurements only in specific time slots, i.e. according to a given sampling ratio. Indeed, they only acquire the required amount of data readings. Relying on the techniques of [21], to compute adaptive compression and sparsifying matrices that vary with the signal correlation structure, we consider a different design that enables us to treat the signal in its matrix form instead of the standard use of the data in CS, i.e. the data vector form. To further improve the network energy savings, the routing is jointly considered with the correlation criteria in the active node selection. Withal, the simulation results shows that we are able to keep a good data reconstruction performance, while reducing significantly the energy consumption. This work was validated in our original paper [22].

In our second contribution, we attempt to address nearly the same issue that is investigated in the first part with the use of different techniques. We propose a data gathering approach based on the Matrix Completion (MC) method, a data sampling and reconstruction technique that, on the heels of CS, has recently emerged. The theory of MC states that if the data matrix has a low-rank or approximately a low-rank structure[4], it can be recovered with high accuracy using the partially received elements [23]. The existence of inactive sensor nodes that do not participate in the data sensing during the entire detection period entails the existence of completely missing data rows in the received data matrix, which unfortunately not only impedes the MC resolution but also pollutes the received data [24]. In this context, we develop a novel structured MC-based framework that guarantees the reconstruction of a significant number of missing data rows thanks to the proposed complementary minimization-based interpolation technique[5]. Furthermore, In order to improve the data reconstruction quality, we propose to perform a sensor nodes clustering phase, so that the participation of the active sensing nodes is scheduled according to the clusters assignment. This preliminary phase is done in order to involve all the detected clusters in the data sensing and avoid disregarding sensor nodes that belong to the small clusters, a deficiency or a slip that can occur with high probability in the purely random data sampling, which is usually used in the conventional MC. This work was validated in our original paper [25].

Although the two previous data gathering schemes provide an efficient solution to reduce the amount of sensed data, minimize the network energy consumption and save its energy, mastering the load balancing between nodes remains a relevant challenge. ***In our third contribution***, we undertake the aforementioned second question issue and propose an energy-aware data gathering strategy aiming to alleviate the uneven

[4] In the context of data matrices, the signal low-rank feature is analogous to the sparsity [19].

[5] The proposed framework is also useful for another challenging scenario; when we have a small number of sensors that have to be deployed in a spacious area. Indeed, either the sensor nodes are costly or the environment is large enough to be content with the limited number of sensors. This may concern also the harsh environments that are difficult to access such as volcanoes and other troublesome environments, where the deployment of many sensor nodes is not practical and becomes expensive. However, in many applications, the amount of gathered data must be significant enough to be processed. The idea here is to place a relatively small number of spatially spaced sensor nodes to control the correlated field under a compression ratio. These sensor nodes represent other sensor nodes that do not really exist.Particularly, the sensory data field is, most of the time, highly correlated and redundant between nearby sensor nodes, which makes possible to estimate readings at locations, where the signal cannot be sensed.

energy depletion problem that may occur in most of the WSNs. The proposed data gathering strategy extends the previous one and selects the representative nodes that can report more information about the others and at the same time afford the sustainability as long as possible for the network lifetime. Since the schemes performance usually vary with the network configurations, we evaluate our approach under different network topologies and scenarios, while selecting, in each time, the adequate energy-aware cost selection function. For each case, the trade-off between the data recovery error and the network lifetime is measured, and the performance behaviour of the proposed data gathering approach is studied for both types of sensor nodes; the low-power nodes and the greedy-power ones (in terms of sensing). This work was validated in our original paper [26].

1.4 Manuscript Organization

This dissertation contains two parts that are divided into 6 chapters. Following this introduction, we discuss some related works in the next chapter. Before going into details, overviews on the CS and MC theories will be introduced. Then, in the second part, we detail the main proposed techniques of this thesis. The subsequent chapters $3-5$ are orderly arranged in accordance with the contributions that have been stated herebefore. More precisely, in chapter 3, we present our routing-aware CS-based approach and describe its components and its design in details, where, differently to most of the existing CS-based schemes, the proposed one integrates both the temporal and the spatial dimensions not only in the data recovery phase but also in the data acquisition one. In chapter 4, we address a challenging compression pattern, which is composed of both structured and random losses, that we successfully manage with the use of a structured MC-based data recovery framework. Chapters 3 and 4 propose also a description of the signals generation models that have been used for the evaluation of the proposed schemes. In chapter 5, we present how the energy constraint can be jointly considered with the correlation criteria in the active node selection cost function in order to maintain a load balancing among nodes and maximize the network lifetime, while still preserving a low data reconstruction error. Finally, we conclude this thesis in chapter 6 by recapitulating our contributions and presenting some eventual perspectives that can be worth pursuing in the future. We present all

the publications of this thesis in Appendix A.

Chapter 2

Backgrounds and related works

Contents

With the emergence and expansion of WSNs applications, reporting a sufficiently accurate description about the monitored environmental phenomena remains of paramount importance. For that reason, investigating and setting efficient data gathering schemes in WSNs have motivated many researchers over the past years.
In this chapter, we start with an overview of the used theories of CS and MC and we discuss some existing data collection schemes from the literature.

2.1 Overview of Compressive Sensing

CS provides a new paradigm that makes possible a high-dimensional sparse signal recovery with the use of a small number of measurements. It is based on two principal conditions: sparsity, which is directly related to the signals of interest, and incoherence, which concerns the data sensing modality [13].

2.1.1 Sparsity Condition

Consider an N-dimensional signal vector $x = [x_1, x_2, ..., x_N]^{tr} \in \mathbb{R}^{N\times 1}$ and suppose that x can be represented in some invertible transformation basis $\Psi = \{\psi_1, \psi_2, ..., \psi_N\} \in \mathbb{R}^{N\times N}$ as:

$$x = \sum_{i=1}^{N} \alpha_i \psi_i = \Psi\alpha, \tag{2.1}$$

where $\alpha = [\alpha_1, \alpha_2, ..., \alpha_N]^{tr}$ holds the transform domain coefficients in Ψ. We say that x is k-sparse in Ψ, if α has at most $k << N$ non-zero entries, i.e. $\| \alpha \|_0 \leq k$, where $\| \alpha \|_0 = | \{i \mid \alpha_i \neq 0, i = 1, ..., N\} |$[1]. In many applications, signals have a few k large coefficients, while the remaining ones are small; in this case we say that x is approximately k-sparse. CS work can be extended to compressible signals which are not exactly sparse. We say that the signal x is compressible if the magnitude of its transform coefficients typically decay according to a power law, that is, $| \alpha_i | < Ri^{-1/p}$ $\forall i$, where $| \alpha_1 | \geq | \alpha_2 | \geq ... \geq | \alpha_N |$, R is a constant, and $0 < p < 1$, i.e, the energy in α is concentrated [27].

[1] $|.|$ presents the cardinality for a discrete set.

2.1.2 Under-sampling and Sparse Signal Recovery

The k-sparse/compressible signal x can be accurately recovered from $M < N$ linear projections ($y \in \mathbb{R}^{M\times 1}$) with high probability [15]. These projections are obtained through an $M \times N$ matrix Φ according to the following equation:

$$\begin{aligned} y &= \Phi.x \\ &= \Phi.\Psi.\alpha \\ &= \Theta.\alpha. \end{aligned} \tag{2.2}$$

Yet, this underdetermined system is ill-posed as the number M of equations is smaller than the number N of unknown variables. Consequently, there exists an infinity of vectors α satisfying (2.2). However, according to the CS theory, if α is sparse or approximately sparse and if the matrix product Θ satisfies the Restricted Isometry Property (RIP) for some isometry constant $0 < \delta_k < 1$[2] [14,28]:

$$(1-\delta_k)\|\alpha\|_2^2 \leqslant \|\Theta\alpha\|_2^2 \leqslant (1+\delta_k)\|\alpha\|_2^2, \tag{2.3}$$

then, it has been shown that recovering the signal x from the projections of y can be achieved through the use of specialized optimization techniques. As an example, we have the Basis Pursuit (BP) convex optimization technique which uses ℓ_1 norm[3] and involves linear programming techniques [15,29,30]:

$$\widehat{\alpha} = arg \min_{\alpha\in\mathbb{R}^{N\times 1}} \|\alpha\|_1 \quad s.t. \quad y = \Theta.\alpha. \tag{2.4}$$

Note that solving (2.2) through the ℓ^1-minimization problem has been adopted as the best alternative convex approximation to the original NP-hard ℓ^0-minimization problem:

$$\widehat{\alpha} = arg \min_{\alpha\in\mathbb{R}^{N\times 1}} \|\alpha\|_0 \quad s.t. \quad y = \Theta.\alpha. \tag{2.5}$$

In the state-of-art, many efficient convex relaxation and greedy pursuit-based solvers have been proposed such L1-MAGIC [30] and Orthogonal Matching Pursuit (OMP) [31].

[2] Broadly speaking, we loosely denote that a matrix Θ obeys the RIP of order k if δ_k is not too close to 1 [13].

[3] The norm ℓ_1 of a vector $x \in \mathbb{R}^{N\times 1}$ is defined by $\| x \|_1 = \sum_{i=1}^{N} | x_i |$, whereas, its ℓ_2 norm is defined by $\| x \|_2 = \sqrt{\sum_{i=1}^{N} x_i^2}$.

Similarly, we can directly minimize the ℓ_0 norm using the Smoothed ℓ^0 (SL0) [32]. In [16, Table. 1], authors have provided details on the different classes of many existing CS recovery algorithms.
Finally, once $\widehat{\alpha}$ is estimated, (2.1) is used to compute the signal $\widehat{x}$. Figure 2.1 provides an illustrative schematic representation of the CS method.

In the case of noisy data, we take into account the additive noise in the obtained measurements, and we replace (2.2) by (2.6), as follows:

$$\begin{aligned} y &= \Phi.x + no \\ &= \Theta.\alpha + no, \end{aligned} \tag{2.6}$$

where $no \in \mathbb{R}^{M\times 1}$ is a vector representing the noise. In most cases, it is considered to be a Gaussian white noise with a zero mean and a variance σ^2_{no}. To approximate the noisy version of (2.5) and search for the sparsest solution $\widehat{\alpha}$ that is consistent with the known or received measurements y, instead of (2.4), we solve (2.7):

$$\widehat{\alpha} = arg \min_{\alpha\in\mathbb{R}^{N\times 1}} \|\alpha\|_1 \quad s.t. \quad \|y - \Theta.\alpha\|_2 \leqslant \varepsilon, \tag{2.7}$$

where ε is an upper bound of the noise [33].

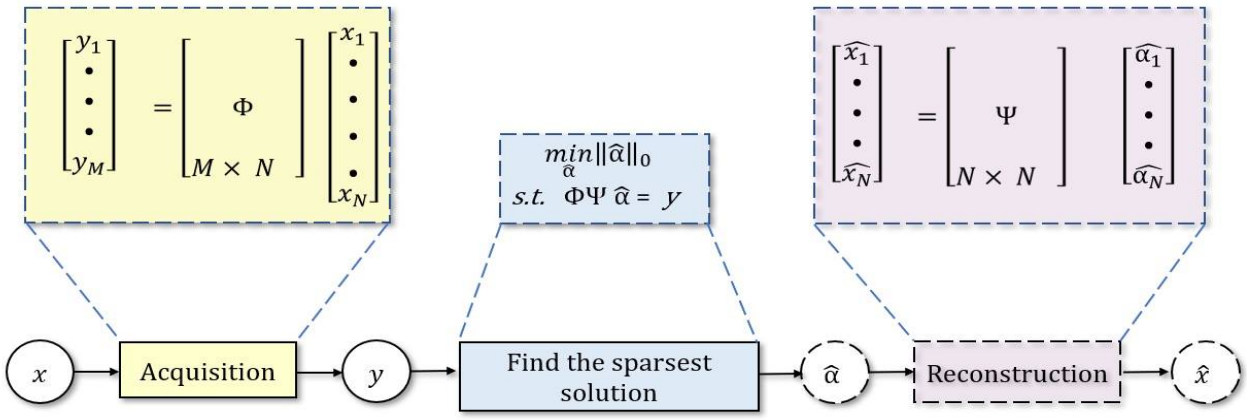


Figure 2.1: Data under-sampling and recovery using CS thechnique.

2.1.3 Incoherence Condition

To achieve a successful CS reconstruction of the signal, another condition must be satisfied, that is, the mutual coherence $\mu(\Theta)$ between Ψ and Φ is required to be small:

$$\mu(\Theta) = \max_{i \neq j\,,\, 1 \leqslant i\,,\, j \leqslant N} \frac{|\langle \theta_i, \theta_j \rangle|}{\|\theta_i\|_2 \|\theta_j\|_2}. \tag{2.8}$$

In this equation, θ_i and θ_j denote the columns of Θ. The mutual coherence $\mu(\Theta)$ determines the number of required projections for an accurate recovery. According to [13], unlike the signal of interest x, the mutual incoherence property (MIP) means that the sensing/compression matrix Φ holds an extremely dense representation in the basis Ψ, and the smaller the coherence (2.8) is, the fewer measurements are required. Differently, the RIP of Θ ensures the measurements or projections to approximately preserve the Euclidean length of all k-sparse signals [34][4]. Since it is difficult to check whether a matrix satisfies the RIP or not, in practice, it is replaced by the mutual incoherence property as shown in [31]: The MIP implies the RIP but the reverse is not true.

Interestingly, the independent and identically distributed (i.i.d) Gaussian and Bernoulli (random ± 1) Φ exhibit a very low coherence with any given orthonormal basis Ψ then satisfies the RIP and ensures an exact data reconstruction with overwhelming probability, if $M \geq C_0.k.\log(N/k)$, where C_0 is a small positive constant [13,35]. Typically, in practice $M = c.k$ with $c \approx 3$ or 4 can be sufficient to meet this condition [18]. However, these dense random matrices[5] still cause high inter-communication costs between sensors and thus limit the efficiency for the applications of CS in WSNs. To overcome such limitations, [27,36]-[38] proposed to use sparse random matrices that contain very few non-zero elements but require a multi-hop routing algorithm establishment. Besides affecting the data recovery performance, the encoding matrix Φ determinates the data readings gathering structure. For that reason, a notable attention has been paid to the design structure of Φ [39,40].

[4] Namely, we verify whether the matrix Θ preserves the distances between all the k-sparse signals, i.e. if the matrix Θ satisfies the RIP, then the distance between two measurement vectors $y_1 = \Theta\alpha_1$ and $y_2 = \Theta\alpha_2$ is proportional to the distance between α_1 and α_2 [20, Chapter. 2].

[5] The use of dense encoding matrices refers to the digital CS [17].

2.2 Overview of Matrix Completion

2.2.1 Under-sampling and Low-rank Matrix Recovery

As an extension of CS, MC technique has emerged recently to benefit from the signal low-rank feature in order to recover the missing data from a substantially limited number of matrix entries [23]. That is, a partially unknown matrix $M \in \mathbb{R}^{N\times T}$ of rank $r \ll min\{N,T\}$ can be entirely reconstructed if a subset of its sampled elements M_{ij} as well as their indices $(i,j) \in \Omega$ are available at the receiver side. The entry-wise partial observation operator $P_\Omega : \mathbb{R}^{N\times T} \rightarrow \mathbb{R}^{N\times T}$ is defined by the following expression:

$$[P_\Omega(X)]_{ij} = \begin{cases} X_{ij} & (i,j) \in \Omega \\ 0 & \text{otherwise.} \end{cases} \tag{2.9}$$

Note that the significance or indication of the notation M in the MC theory differs from that in the CS theory. Here, M refers to the received data matrix to be recovered, whereas, in the CS theory, M refers to the number of received measurements that compose the projection vector y (2.2).

Roughly speaking, the goal of the MC is to find a low-rank matrix X that is consistent with the observed measurements M_{ij}. According to [23], if Ω contains enough information and if $M \in \mathbb{R}^{N\times T}$ is a low rank or approximately a low-rank matrix, we can fill the unknown entries by solving the following rank minimization problem:

$$\underset{X\in\mathbb{R}^{N\times T}}{\text{minimize}} \; rank(X) \quad s.t \quad P_\Omega(X) = P_\Omega(M). \tag{2.10}$$

Yet, problem (2.10) is not convex, and algorithms solving it are doubly exponential. Fortunately, the nuclear norm $\| X \|_*$ minimization problem, which is a convex relaxation, can be solved. In fact, it is deployed as an alternative to the NP-hard rank minimization problem [41]. Thus, we have:

$$\underset{X\in\mathbb{R}^{N\times T}}{\text{minimize}} \; \| X \|_* = \sum_{i=1}^{r} \tau_i(X) \quad s.t \quad P_\Omega(X) = P_\Omega(M). \tag{2.11}$$

$\| X \|_*$ denotes the sum of the singular values $\tau_i \geq 0$ of the matrix X. As it might be seen, the relationship between the nuclear norm and the rank function in MC is analogous to that between the convex ℓ_1 norm and the ℓ_0 norm in CS. Indeed, while

the rank provides the number of non zero singular values $\tau_i > 0$, the nuclear norm measures their sum.

In the literature, various efficient solvers for this type of systems have been suggested. For example, the Singular Value Thresholding (SVT) optimizes an approximation of (2.11) by using a threshold parameter τ_{au} and adding a Frobenius-norm term to the objective function [42]:

$$\underset{X \in \mathbb{R}^{N \times T}}{\text{minimize}} \ \tau_{au} \parallel X \parallel_* + \frac{1}{2} \parallel X \parallel_F^2 \quad s.t \quad P_\Omega(X) = P_\Omega(M). \tag{2.12}$$

Different from (2.11), another method has been proposed to approximate (2.10) rather than the nuclear norm, which is the matrix factorization. Low rank matrix fitting (LMaFit) [43], Sparsity Regularized SVD (SRSVD) and Sparsity Regularized Matrix Factorization (SRMF) [44] are among the approaches that use the matrix factorization method. These approaches are based on the fact that any matrix $X \in \mathbb{R}^{N \times T}$ of a rank up to r can be explicitly written as the product of two matrices with the form $X = LR^{tr}$, where $L \in \mathbb{R}^{N \times r}$ and $R \in \mathbb{R}^{T \times r}$. Hence, the goal here is to search over the set of rank-r matrices and find a point LR^{tr} that is closest to the set of matrices, which meets M at all known entries. To solve the problem, an alternating minimization scheme is used by fixing one of L and R and making the other one as the optimization variable.

2.2.2 Incoherence Condition

As with the CS theory, from a theoretical point of view, in order to find the desired solution with this kind of methods, the sampling set Ω must be selected uniformly at random[6]. However, it has been shown in [23] that it is impossible to get that kind of guarantees of the MC-based recovery to all the low-rank matrices. To see the problem, suppose that the rank-r singular value decomposition (SVD) of the known data matrix M is $U \Pi V^{tr}$, where $V \in \mathbb{R}^{T \times T}$ and $U \in \mathbb{R}^{N \times N}$ are two unitary matrices. Besides, we assume that $r = 1$ and both (or one) singular vectors are sparse, i.e. their total energy is carried only by few entries. Yet, when this occurs, the resulting matrix M will, as well, hold its energy concentrated on just a few number of its entries, i.e. M equals to

[6] From the matrix-RIP theory point of view, we verify if the operator P_Ω preserves or not the distances between all the rank-r data matrices [45] [46].

zero in almost all columns or rows [45]. In such particular situations, it is impossible to find M unless all of its elements are observed[7]. This example illustrates that one cannot hope to fill or complete the data matrix if some of the singular vectors are extremely sparse[8] [41]. To avoid such informal considerations and particular situation, the singular vectors of M should be spread across all the coordinates. The authors of [23] have introduced a geometric incoherence assumption, that is, M has to satisfy the incoherence condition with parameter μ_0 as follows:

$$\begin{aligned} \max_{1\leqslant i\leqslant N} \parallel U^{tr} e_i \parallel_2 \leq \sqrt{\frac{\mu_0 r}{N}}, \\ \max_{1\leqslant j\leqslant T} \parallel V^{tr} e_j \parallel_2 \leq \sqrt{\frac{\mu_0 r}{T}}, \end{aligned} \tag{2.13}$$

where $\{e_i\}$ and $\{e_j\}$ both represent the canonical basis for the appropriate dimension and $1 \leq \mu_0 \leq \frac{min\{N,T\}}{r}$. Fortunately, this is usually the case in most of the practical applications. According to [23], most matrices M of low rank r can be perfectly recovered with probability $1 - n_c^{-3}$, and the solution of (2.11) will converge to the solution of (2.10), if the number of received data samples is in the order of $m_M \geq C_c n_c^{6/5} r log(n_c)$, where C_c is a constant and $n_c = max(N, T)$ [47, 48].

2.3 CS and MC based data gathering approaches

Environmental WSN signal profiles exhibit both spatial and temporal dependency. Such structures generate redundancy and enable a succinct representation of the data using a number of coefficients much smaller than its actual dimension. One popular postulate of such low-dimensional structures is sparsity, that is, a signal can be simply represented with a few non-zero coefficients in an invertible proper sparsifying domain [49]. With a number of measurements proportional to the sparsity level, CS enables a reliable reconstruction of the signal. Over the past years, plenty of papers have addressed the data gathering problems in WSNs by the integration of the CS theory to drastically reduce the number of transmitted measurements [21,27,36]-[38,50].

[7] An analogous situation with the CS is that one evidently is unable to reconstruct a signal, which is sparse in the time domain, through sub-sampling it in the time domain.

[8] Generally, in the case where a column (or row) does not have a relationship to the other columns (or rows) in such a way that they are approximately orthogonals, basically we would require to observe all the data entries in that column (or row) to reconstruct M.

Originally, CS-based schemes were designed to sample and recover sparse vectors and were classified either as purely spatial approaches [27,36,37,50,51] or as purely temporal ones [52]. For the spatially (inter-sensors) CS-based data gathering approaches, in each time slot, along the multi-hop path that relays the initial transmitting source node to the sink, the readings of the relaying sensor nodes and the initial transmitting source node are linearly combined using their coefficients of Φ, resulting to a measurement projection of a weighted sum. Here, each vector row of the compression/measurement matrix Φ represents a path and holds non-zero coefficients only in the positions of the initial sensor node and the relaying ones. Given the example of Figure 2.2, suppose that the applied compression ratio imposes the collection of $M = 2$ measurements projection, i.e. $y = [y_1, y_2]^{tr} \in \mathbb{R}^{M \times 1}$. In this case, each of the initial transmitting source nodes N_2 and N_5 initiates a separate projection that is computed hop by hop until being received by the sink. However, this kind of in-network aggregation scheme is highly dependent to the considered routing rules and to the network topology [36,37,53].

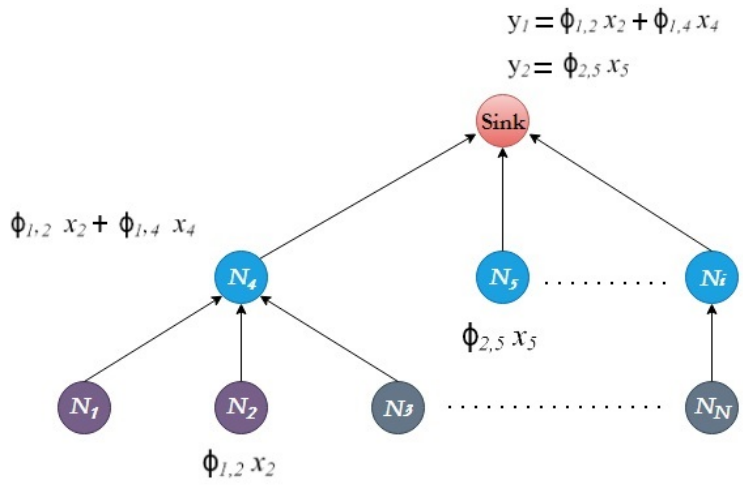


Figure 2.2: Example 1 of data gathering process using spatial CS method.

As another example of typical application of the spatial CS, illustrated in Figure 2.3, all the leaf nodes initiate the data transmission process. For each projecton y_j, sensor node N_i multiplies its probed data reading x_i by its coefficient $\phi_{j,i}$. The resulting partial projetion is added to the received ones that are computed by the children nodes and then forwarded to the higher node. Even though, this method implies the use of

dense encoding matrices Φ, compared to the baseline data collection, this kind of CS data gathering scheme reduces the number of messages to be delivered to the sink for large-scale WSNs, i.e. when M is much smaller than N and N is too large. Moreover and more importantly, the transmission load is uniformly spread out between all nodes since they forward the same size of information whatever the distance from the destination is [50].

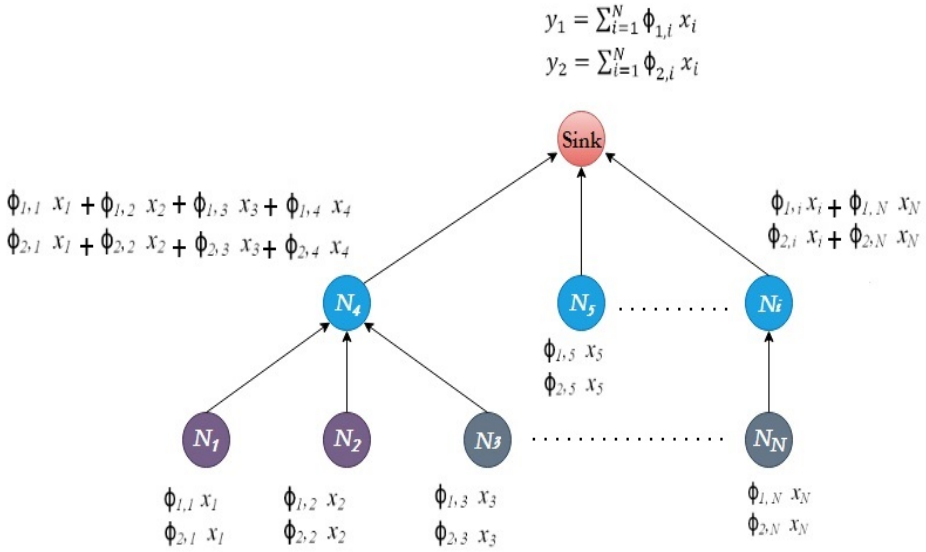


Figure 2.3: Example 2 of data gathering process using spatial CS method.

For the temporally (intra-sensor) CS-based data gathering approaches, each sensor node reports to the sink only the CS measurements projection, obtained from a block of its data readings that are sampled during a number of successive time slots then buffered [52, 54]. Different from the spatial CS methods, this in-node compression technique is localized and network independent. To this end, chapter 3 relies on the idea of exploiting both the distributed (spatial) and local (temporal) CS designs to deliver only a fraction of data sensory readings to the sink without any on-board sensor nodes computation[9].

The inherent correlation between sensory data readings enables the data, probed by nodes during a period of time, to exhibit a low rank structure, which is analogous

[9]Note that a more detailed discussion of how CS methods have been applied in WSNs is afforded in the introduction of the chapter 3.

to sparsity. Following the CS, the MC theory presents a remarkable new field that takes advantage of the low-rank feature of the data matrix to recover the missing entries. In [55], a state-of-the-art of MC-based algorithm for compressive data gathering has introduced the short-term stability with the low-rank feature. The considered feature was used not only to reduce the recovery error but also to recover the likely empty columns appearing in the received data matrix. The existence of the empty columns was possible since the readings were forwarded according to a presence probability. Furthermore, authors in [56] addressed joint CS and MC. They used the CS to compress the sensor node readings then the MC to recover the non-sampled or lost information. However, this approach has not been compared to other state-of-the-art approaches to show its real contribution. In addition, they didn't take advantage of the space-time correlation of the signal as it should be, since they used standard compression and sparsifying matrices for the CS. Different from [56], Wang et al., in [57], explored the graph based transform sparsity of the sensed data and considered it as a penalty term in the resolution of the MC problem. Similarly, [58] has combined the sparsity and the low-rank feature in the decoding part and, as in [57], used the alternating direction method of multipliers to solve the constrained optimization problem. Since adaptability and efficiency are two very important issues in WSNs data gathering, [59] proposed an adaptive and online data gathering scheme for weather data that is purely based on the MC requirements. Yet, the main drawback of this approach was the computational overhead at the sink to reconstruct and re-reconstruct the same active window data as well as the extra communication cost between the sink and the sensor nodes in order to adjust the number of needed measurements. The process is reiterated until the required error gap is reached, even though they have already found a very low reconstruction error. In contrast to our proposed approaches, this paper addressed the sampling side. Indeed, they focused on the sampled data locations in the received data matrix, whereas, we have considered the sampled data locations in the network area. Authors of [24] focused on the case of MC recovery with the existence of successive data missing or corruption, referred to as structure faults. Indeed, they considered that successive data may be missing or corrupted due to channel fading or sensor node failures, which creates successive missing data on rows and/or on columns. Although this successive missing data may result in the existence of some few empty rows, the proposed data reconstruction approach does not take into account these particular totally empty rows, and fails to recover the data matrix

when the number of missing rows becomes significant. Hence, in our work of chapter 4, we investigate how to solve a challenging problem in the WSNs: how to omit a considerable number of sensor nodes from the monitoring schedule and estimate their readings from the partially reported readings of a set of representative sensor nodes using a MC-based approach.

2.4 Energy-efficient based data gathering approaches

In the state-of-art of the energy-efficient based algorithms for data gathering, reducing the amount of collected data readings or reducing the packets size are two well investigated methods that are closely related to the minimization of the network energy consumption [12, 21]. CS and MC take benefits from the redundancy that occurs in the environmental WSN signals in order to reduce the number of transmitted measurements and thus achieve an appealing progress in the network energy consumption [7, 24, 60]-[62].

Li et al., in [62], have combined the CS and the routing scheme and proposed a multi-strip data gathering approach for green data collecting. Using this approach, the network is partitioned into multiple strips, where nodes around each strip forward data to the center with data fusion technique. The amount of data readings undertaken by sensors is relatively balanced since the transmitting nodes are changing. Yet, according to [63], this scheme doesn't use an adaptive distributed technique to minimize the complexity in data gathering. On the heels of MC, Tan et al., in [61], targeted to enhance the network energy efficiency and proposed a low redundancy data collection scheme. This MC-based approach serves to quickly compensate the set of collected data in cases of packet loss. In order to not affect the network lifetime, this approach takes advantage of the energy surplus, remaining away from the sink area, and conceives the backup data set to satisfy the minimum number of measurements required by the MC theory. Different from the compression-based aspect that the aforementioned schemes have proceeded, the authors of [64] have addressed the network lifetime issue by reducing the number of nodes' state transitions, pointing out that the processor consumes energy through state transition. This technique bears a resemblance to ours in the sense that, in our scenario, a set of sensor nodes is scheduled to not sense the environment for a large number of consecutive time slots in order to

reduce their power consumption.

In line with the consideration of the transmission path to increase the network lifetime, Yao et al., in [65], have developed an energy-efficient delay-aware lifetime-balancing data collection algorithm for heterogeneous WSNs, in which nodes holding poor communication links and less remaining energy have a lower chance to be chosen as forwarders. At the beginning of each collection period, a set of nodes is selected to be the sources. However, in our proposed approaches of chapter 4 and chapter 5, the source nodes differ from a time slot to another ensuring a diversity in the reported data and thus a better monitoring quality and energy balancing. Similarly, the paper [66] has proposed two algorithms, where a sensor node always chooses, as next hop, the node that has the highest residual energy. Yet, the proposed techniques have been proved in [67] to be unable to manage the problem of void hole. To overcome the energy hole problem, authors in [68] have introduced a new layer, referred to as the charging layer, into the basic node network protocol stack. As soon as the battery level of a node goes done, it is charged wirelessly using witricity (wireless electricity). In this context, the EHWSNs, where nodes can replenish their batteries with energy from the surrounding environment, have got attention of several researchers. Among the in-network processing-based schemes, an m-hop averaging data compression technique, with energy harvesting, has been proposed in [12] in order to deal with the unevenness of the energy levels among the nodes. In this algorithm, each node has to continuously assemble the usable energy levels of other nodes then make a decision about how much it needs to compress the forwarded data packets after comparing its own energy level with those assembled from the next m nodes within m hops. As the packet is relayed towards the sink, the data packet length becomes smaller leading to a gradual decrease in the energy cost. Different from [12], in [60], authors have presented an adaptive collection scheme-based MC, which adjusts the amount of data to be gathered at each moment depending on the residual usable energy absorbed from solar radiation. This scheme has been designed to improve the network energy utilization, increase the duty cycle of sensors far away from the sink and gather as less data readings as possible, when there is no sufficient usable energy and vice versa. Yet, this is not the case with our scenario since our deployed sensor nodes can neither charge their batteries nor renew them. Dealing only with the nodes' batteries, we have investigated in chapter 5 tow to extend the network lifetime and prevent it from

being prematurely partitioned or dead by considering the residual energy of the entire multi-hop path that links the source node with the sink. Furthermore, at the end of the network lifetime, the remaining energy of the border nodes (i.e. nodes far away from the sink) is almost close to the average remaining energy thanks to the introduced energy-aware cost functions that select the representative sensor nodes. Without any extra communication between nodes, the proposed metrics aim not only to achieve energy efficiency but also to preserve a sufficiently good quality of data reconstruction as they take into account correlation among sensors to select those who can report more information about the network.

Part II

Compressive Sensing and Matrix Completion based approaches in Wireless Sensor Networks

Chapter 3

Space-Time Compressive Sensing Routing-Aware approach

Contents

3.1 Introduction

Recently, it has been shown that the incorporation of CS techniques has enhanced WSNs scenarios since they have been introduced as a good fit for such applications in both, the acquisition as well as in the reconstruction of the signal [69].

To achieve a successful application of CS in WSNs, incoherence condition between the transformation matrix Ψ and the measurement matrix Φ must be present while simultaneously considering data gathering problems and communication cost. In this context, [36] and [27][1] addressed the impact of the network topology and the routing system on the CS process in WSNs. [36] found that none of the standard transformations can sparsify the signal in question while being simultaneously incoherent with the measurement matrix Φ, which badly affects the recovery performance. [27] presented a centralized algorithm that iteratively build projections and choose paths that minimize the intermediate coherence with a given Ψ in order to reduce the reconstruction error. However, no performance improvement was found compared to the randomized downsampling. Likewise, [37] studied the problem of data gathering using CS in WSNs and graph theory. They provided mathematical foundations for a novel approach leading to a non-uniform collection of measurements through a random walk based manner. Yet, the problem with this approach is the direct influence of the random selected paths on the compression performance. As a sequel of [36], Quer *et al.*, in [51], came with the idea of the online estimation of Ψ, exploiting the Principal Component Analysis (PCA) approach, referred by many authors as Karhunen-Loève Transform (KLT), to capture the temporal or the spatial characteristics of the received signal. Basically, the idea of the PCA technique is to rotate the axes of the data in order to minimize the correlation that can be interpreted as redundancy between coefficients and as a result increase the energy concentration. Particularly, the basis vectors of the PCA matrix are given by the orthonormalized eigenvectors of the data autocorrelation or covariance matrix [70]. The results of paper [51] have attracted the attention to the CS when it is used as a recovery tool in WSNs. [21] and [52] used also the PCA technique by making adjustments according to their applications. Hooshmand *et al.*, in [21], added the covariogram computation to the standard PCA to get a better estimation of the spatial transformation matrix. On the other hand,

[1] They are two state-of-the-art studies for the CS-based approaches in WSNs.

Chen *et al.*, in [52], used the incremental PCA to calculate the temporal dictionary, which stores in memory just the k largest eigenvalues of the covariance matrix.

Since environmental WSN signals have, most of the time, both temporal and spatial dependency, this characterization was then exploited by the incorporation of the kronecker CS framework [21, 28, 34, 71]. However, in [34] and [21] the integration of the multi-dimensional CS aspect was done on the sparsifying level, ignoring the compression one which is highly important in the case of WSNs. Wang *et al.*, in [71], proposed a $2D$ data gathering strategy called CS²-collector, which applies CS locally at each sensor as well as in the whole network. However, the proposed approach didn't take advantage of the $2D$-correlation existing in the signal as it should be, since it uses standard transformation matrices. These data independent matrices ignore how the signal is correlated and when its correlation changes, which leads to a limited reconstruction performance. Inspired by the CS²-collector model, and relying on the CS mechanisms used for the CB-CS (Covariogram-Based Compressive Sensing [21]), we include in our design, the temporal sampling and sparsifying pattern to compress, then, reconstruct the signal in an efficient way through the Space-Time Compressive Sensing (STCS) approach.

WSNs possess a finite and limited power supply capacity [72]. For that reason, the primary factor to consider is the minimization of the energy consumption, even though this may affect or degrade a little bit the recovery performance. As the CS approach is based on transmitting a small number of coefficients rather than the full set of the signal coefficients, it provides schemes that can reduce efficiently the network power consumption, as shown in [7, 73, 74]. In this direction, to further increase the network energy saving, we integrate the routing in the active node selection process through the STCS-Routing Aware (STCS-RA). Several researches used the routing in conjunction with compression in order to linearly combine sensors readings along the multi-hop selected paths [27, 36]-[38]. However, in this work, the routing is used in conjunction with the spatial correlation in order to select the nodes that can best present the whole network, when at the same time, are "near" the sink.
The main contributions of this chapter are summarized as follows:

- In the data gathering part, only a small subset of sensor nodes is selected to be active and report their readings to the sink. These sensor nodes should capture

enough information to be chosen as the representative of the network. In the following, we define the node selection criterion that allows the sink to recover the entire data. Correlation among sensors is calculated and those holding the greatest informative values are better ranked to be chosen.

- Both distributed and local data gathering based on CS technique are efficiently investigated in this work. Different from [71], the temporal compression pattern of our approach with its sparse combination does not entail on-board sensors computation. Making use of this kind of conception for the compression matrices meets well the constraint of limited computational capacities that characterizes the sensor devices. Consequently, the proposed STCS consumes less energy than [71], while reaching higher data recovery quality.
- If the gathered data is expressed in the vector form, as it is usually the case in the standard CS, spatial and temporal correlations can not be handled together. Thus, to take benefits from both inter and intra-dependency, the signal is treated in its $2D$ form, using tools from linear algebra.
- Finally, the sensor route length is taken into account with the STCS-RA in the active node selection phase in order to significantly improve the trade-off between minimizing the energy consumption of the network and maintaining a good reconstruction quality.

This chapter is organized as follows. Section 3.2 defines the network model and the signal model. Section 3.3 presents the proposed algorithm and describes its components and its design in details. In Section 3.4, we carry out with simulations to show the performance of our STCS and STCS-RA. Finally we conclude the chapter in section 3.5.

3.2 System Model

3.2.1 Network Model

We consider a multi-hop wireless sensor network consisting of a set $\mathcal{N}_f = \{1, \ldots, \mathrm{N}\}$ of randomly distributed sensors in a square observation area. We assume that the sink is located at the center of the area to gather the transmitted measurements, and we

suppose that it has an infinite power supply. We consider that two nodes are connected only if the Euclidean distance between them is shorter than some transmission radius (r) that scales with $\ominus(\sqrt{logN/N})$ to guarantee the connectivity of the network with high probability [37, 75].

To route the data towards the sink we use the shortest path tree computed by Dijkstra algorithm [76][2]. Figure 3.1 includes an example of a routing tree, found by Dijkstra algorithm, for a network composed of $N = 50$ sensor nodes.

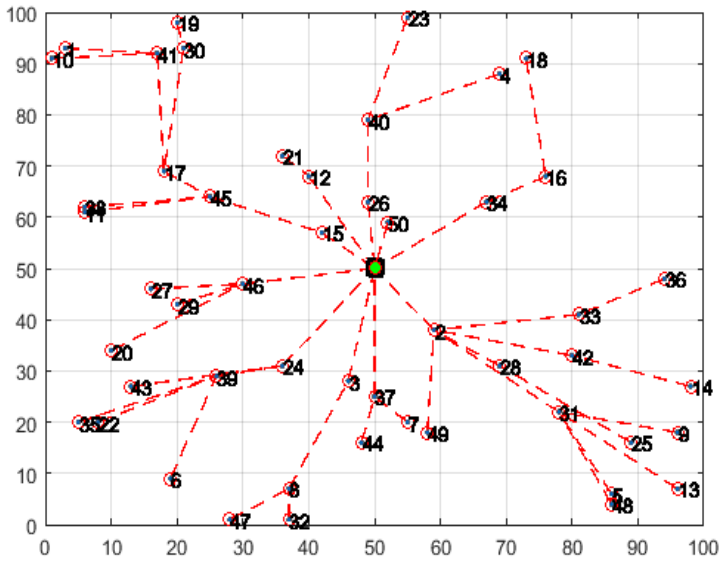


Figure 3.1: A routing tree for a network composed of $N = 50$ sensor nodes.

3.2.2 Signal Model

Since the temporal and the spatial correlations represent a huge potential for compressing especially in the WSN signal profile, it will be interesting, if we can vary numerically their degrees to figure out how would be the performance of any proposed compression scheme [21]. This solution was introduced by [77], which allows the efficient generation of a synthetic continuous space-time signal field, where spatial

[2] According to [76], paths established by Dijkstra algorithm usually present a lower number of connections, hence, the average delay of message dissemination decreases which reinforces the energy management.

($\gamma > 0$) and temporal ($\rho \in [0, 1]$) correlation parameters can be separately adjusted[3], since their corresponding functions are independent.

To generate the signal of interest, we suppose that $D = [-x_D, x_D] \times [-y_D, y_D]$ is the space domain, where x and y are the space coordinates. Moreover, we suppose that the time is slotted into equal time slots $t = 1, 2, \ldots, T_{cs}$. Algorithm 1 states how to generate a correlated stationary signal field $z(p, t) : D \times \mathcal{T} \rightarrow \mathbb{R}$, where $\mathcal{T}$ is the time domain and p is a point in (x, y) plan D.

To start the signal generation process, for $t = 1$, we define $w(p, t) : D \times \mathcal{T} \rightarrow \mathbb{R}$ to be an i.i.d random Gaussian field. More precisely, for any specific position $p = (x, y)$, $w(p, 1)$ is a Gaussian random variable with zero mean and unit variance.

To obtain a temporally correlated signal, authors of [77] have used an autoregressive filter to enforce the temporal correlation in the signal model (step 3 of the algorithm 1). Since the time is slotted into equal time slots, they only consider the one-step time correlation and use a simple coefficient ρ. Note that it has been shown in [77, Eq. 8 and Eq. 9] that the performed autoregressive model maintains the statistical properties and preserves the mean and the variance of the initial used signal $w(p, 1)$, i.e $\mu_{w(p,t)} = 0$ and $\sigma^2_{w(p,t)} = 1$, $\forall t \in \mathcal{T}$.

Regarding the spatial correlation, we apply to the signal, to be generated, a $2D$ filtering procedure using a specific correlation function $rs(p)$ (step 6 of the algorithm 1). Among the numerous existing models in the literature, we generate the signal using the Gaussian filtering[4], used in [21, Eq. 2], which can be controlled by the parameter γ:

$$rs(p) = exp(\frac{-(x^2 + y^2)}{\gamma \alpha_s}). \tag{3.1}$$

In (3.1), α_s is a scaling parameter that depends on the size of the field. In [77], authors stated that the coloration of the signal with $rs(p)$ has to be done in the frequency domain. Hence, before modeling the spatial correlation, a Fourier transformation is performed (step 5 of the algorithm 1). Note that it has been proven in [77, Eq. 12] that the signal field z is still stationary and Gaussian with zero mean ($\mu_{z(p,t)} = 0$).

[3] The values of γ and ρ are in the same order of magnitude as those of the empirical values found in [77].

[4] The Power Exponential model, when ν is equal to 2 [78].

Algorithm 1 Model for generating the correlated signal field.

Input: the generated field for $t = 1 : w(p,t)$, the temporal correlation parameter ρ, the spatial correlation parameter γ, the spatial correlation function computed in the frequency domain $Rs(\omega) = F(rs(p))$.

1: **for** $t = 1$ to T_{cs} **do**
2: **if** $(t \neq 1)$ **then**
3: $w(p,t) = \rho \times w(p,t-1) + \sqrt{1-\rho^2} \times \varepsilon(p,t)$, where $\varepsilon(p,t)$ is a $\mathcal{N}(0,1)$ i.i.d random Gaussian noise.
4: **end if**
5: $W(\omega,t) = F(w(p,t))$.
6: $Z(\omega,t) = W(\omega,t) \times Rs(\omega)^{1/2}$.
7: $z(p,t) = F^{-1}(Z(\omega,t))$.
8: **end for**

Output: the space–time correlated signal field $z(p,t)$.

By construction, the signal field $z(p,t)$ is a $3D$ matrix of size $(2y_D \times 2x_D \times T_{cs})$. The data matrix of interest, $X_{cs} \in \mathbb{R}^{N \times T_{cs}}$, denotes the $2D$ signal discretized from $z(p,t)$ by the N sensor nodes along the T_{cs} time slots, where the $(i,t)^{th}$ entry of X_{cs}, $x_{cs_{i,t}}$, represents the t^{th} data reading $(t \in [1, T_{cs}])$ sensed by the i^{th} sensor node $(i \in \mathcal{N}_f)$.

3.3 Space-Time Compressive Sensing Routing-Aware approach (STCS-RA)

3.3.1 Space-Time compression matrices

3.3.1.1 Spatial sampling pattern

In this part, we explain how routing can be jointly considered with the correlation criteria for the active sensor selection in order to minimize the overall network consumption. At the beginning of each sensing period, the sink selects a set $\mathcal{M}_s = \{1, ..., M_s < N\}$ of sensors that can best represent the whole network and, at the same time have the shortest path towards the sink. Relying on the the Enhanced Correlation Based Deterministic Node Selection (ECB-DNS) procedure [21], we select the M_s sensor nodes according to their conditional variances, computed through [21, Eq. 11], which help selecting the sensor g^* with the maximum informative value m' respecting to the set

of sensors that are not yet selected: S_1. That is[5]:

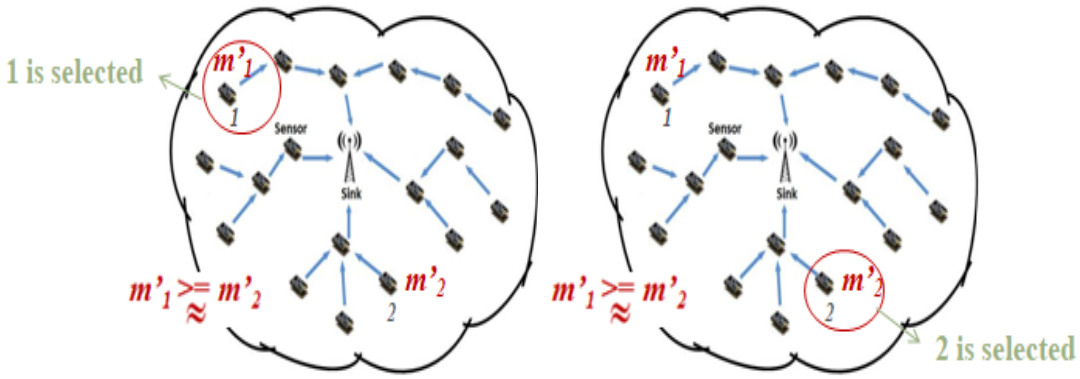
$$g^* = arg \max_{g \in S_1}(m'_g), \tag{3.2}$$

where

$$m'_g = \left(\sum_{i \in S_1} \frac{\sigma_{ig}^2}{\sigma_g^2} \right). \tag{3.3}$$

In equation (3.3), σ_{ig} is the covariance between the variable x_i of node i and the variable x_g of node g and σ_g^2 is the variance of the variable x_g. Differently to [21], in which all the N sensors participate in the transmission along the T_{cs} slots, in STCS, only M_s sensors will be the representatives and transmit for the entire sensing period.

To see the problem when considering only the metric of (3.2), suppose that the selected node is faraway from the sink, while there is another node with slightly the same metric but near the sink. In this case, it would be a waste of energy for the network, if we keep using the selected node. Figure 3.2 illustrates a simplified clarification of the problem. As mentioned above, the node selection process for the STCS-RA takes into



(a) Selecting a sensor node without routing consideration.

(b) Selecting a sensor node with routing consideration.

Figure 3.2: Active sensor node selection.

consideration not only correlation between sensors but also their paths cost, measured

[5]Basically, the overall conditional variance m_g of sensor node g is defined by $m_g = \left(\sum_{i \in S_1} \sigma_i^2 - \sum_{i \in S_1} \frac{\sigma_{ig}^2}{\sigma_g^2} \right)$. Since the first sum does not depend on node g, we rely only on the second sum for reasons of simplicity.

with number of hops. Therefore, we add to (3.2) an additional penalty modeled by the sensors paths costs. For a given sensor $g \in \mathcal{N}_f$, the balance between its m' value of (3.2) and its path cost towards the sink, $nbHops(g)$, is controlled by a tuning parameter β. Thus, (3.2) is replaced by (3.4) for our STCS-RA:

$$g^* = arg \min_{g \in S_1}(-m'_g + \beta.nbHops(g)). \tag{3.4}$$

Generally, the covariance takes its values in the interval [-1, 1]. Thus, the fractions $(\sigma^2_{ig}/\sigma^2_g) \ll 1$. Furthermore, at each selection iteration, these values still decrease until being insignificant, as it will be explained hereafter. Besides, $nbHops$ is an integer $\geqslant$ 1 and it varies according to the transmission radius. Therefore, the values assigned to β must be much less than 1 in order to not neglect the weight of the correlation presented by m'.

For the rest of this chapter, we refer to STCS when we use (3.2) and STCS-RA when we perform (3.4) for the transmitting source nodes selection procedure.

The node selection algorithm is detailed as follows. At the iteration $n \in \{1, \ldots, M_s\}$, a sensor $g^*(n)$ is selected and moved from set S_1 to set S_2. Note that S_2 is the set containing the sensors that are already selected over the previous selection iterations. The metrics m' of the sensors of the set S_1 will be recomputed in order to cancel out the impact of the selected node $g^*(n)$ on the rest of the sensors of S_1 and to prepare for the selection of the next sensor node $g^*(n+1)$. The selection of the node $g^*(n+1)$ will be done as if the node $g^*(n)$ did not exist in the network. The process is reiterated until the selection of $M_s < N$ sensors. The node selection process, especially the manner how we remove the correlation effect of node $g^*(n)$ from S_1, follows the steps outlined in algorithm 2. At the initialization and before the first sensing period, we define the data matrix $X_{lp} = [x^{tr}_{lp\,1}, x^{tr}_{lp\,2}, \ldots, x^{tr}_{lp\,N}]^{tr} \in \mathbb{R}^{N \times T_{lp}}$ that is delivered during a short learning period $T_{lp} \ll T_{cs}$, where all sensor nodes report their information to the sink[6]. We assume that the spatial correlation feature inherent in X_{lp} reflects that in X_{cs}.

Once the best M_s sensors are selected, the compression operation can be represented

[6] $x_{lp\,i} \in \mathbb{R}^{1 \times T_{lp}}$, considered as a T_{lp}-dimensional data points, holds the readings sent by the sensor node i during the learning period.

Algorithm 2 The representative sensor nodes selection process.

Input: $n = 1$, $S_1 = \mathcal{N}_f$, $S_2 = \{\varnothing\}$, $\mathcal{M}_s = \{\varnothing\}$, $X_1 = X_{lp}$, a zero-vector $X_2 \in \mathbb{R}^{1\times T_{lp}}$.

1: **for** $n = 1$ to M_s **do**
2: **if** $(n == 1)$ **then**
3: Compute the covariance matrix $\Sigma \in \mathbb{R}^{N\times N}$ of X_{lp}.
4: According to (3.3) and using Σ, compute the metrics m'. Then, select $g^*(n)$ using (3.2) or (3.4).
5: Remove the reading $x_{lp\,g^*(n)}$ of node $g^*(n)$ from X_1 so that it becomes $X_1 = [x^{tr}_{lp\,1}, x^{tr}_{lp\,2}, \ldots, x^{tr}_{lp\,g^*(n)-1}, x^{tr}_{lp\,g^*(n)+1}, ..., x^{tr}_{lp\,N}]^{tr} \in \mathbb{R}^{N-n\times T_{lp}}$ and X_2 takes the values of node $g^*(n)$ so that $X_2 = x_{lp\,g^*(n)}$.
6: Following that removal, Σ can be written as:

$$\Sigma = \begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{bmatrix}, \tag{3.5}$$

where $\Sigma_{1,1} \in \mathbb{R}^{N-n\times N-n}$ is the covariance matrix of X_1, $\Sigma_{1,2} = \Sigma^{tr}_{2,1} \in \mathbb{R}^{N-n\times 1}$ is the covariance vector between X_2 and X_1, and $\Sigma_{2,2}$ is the variance of X_2.
7: **else if** $(n \geq 2)$ **then**
8: Following the removal of node $g^*(n-1)$ from S_1, re-compute the conditional covariance matrix of X_1 knowing $X_2 = x_{lp\,g^*(n-1)}$; $\Sigma_{1,1|2} \in \mathbb{R}^{N-(n-1)\times N-(n-1)}$, where:

$$\Sigma_{1,1|2} = \Sigma_{1,1} - \Sigma_{1,2}(\Sigma_{2,2})^{-1}\Sigma_{2,1}. \tag{3.6}$$

9: According to (3.3) and using $\Sigma_{1,1|2}$, re-compute the metrics m'. Then, select $g^*(n)$ using (3.2) or (3.4).
10: Σ takes the values of $\Sigma_{1,1|2}$.
11: Perform step 5 then step 6.
12: **end if**
13: $S_1 = \mathcal{N}_f \setminus \{g^*(n)\}$ and $g^*(n) \in S_2$.
14: **end for**

Output: $\mathcal{M}_s = S_2$.

by a left multiplication of the $2D$ signal X_{cs} with the spatial projection matrix $\Phi_S \in \mathbb{R}^{M_s \times N}$ as:

$$Y_S = \Phi_S . X_{cs} + No, \tag{3.7}$$

where $Y_S \in \mathbb{R}^{M_s \times T_{cs}}$ is the spatially compressed $2D$ signal, $No \in \mathbb{R}^{M_s \times T_{cs}}$ is the measurement noise and Φ_S is a sparse matrix that consists of a single '1' in each row and at most a single '1' in each column[7]. Note that '1' corresponds to a selected sensor node $i_{Ms} \in \mathcal{M}_s$, which means that exactly M_s nodes are selected as sampling nodes and they cannot transmit their readings twice. According to [13], this kind of data sampling is denoted as the Nonuniform Sampler (NUS). Here, the i^{th}_{Ms} row of Y_S holds the set of T_{cs} data readings, sensed by the i^{th}_{Ms} sensor node belonging to $\mathcal{M}_s$. The columns of the matrix Φ_S are thus orthogonal, and according to [27], Φ_S is a valid compression matrix[8].

As a result, in contrast to [21], where they have to generate a different compression matrix for each time slot t to compress separately the columns of Xcs, our approach needs just one Φ_S to compress the entire data matrix Xcs.

3.3.1.2 Temporal sampling pattern

Different from [21], for an entire sensing period, the selected sensors are the same. This allows us to compress also the timing signal at each sensor. Thus, each of the selected M_s sensors applies the CS locally on its temporal data vector in order to reduce its dimension, from T_{cs} to $M_t < T_{cs}$ readings, using a sparse random sampling pattern. This can be easily implemented by sharing the seed of random generator. For example, the sink can broadcast the seed to all the sensors at the beginning of each sensing period. These operations can be represented by a right multiplication of the matrix Y_S with the temporal projection matrix $\Phi_T \in \mathbb{R}^{T_{cs} \times M_t}$ as:

$$Y = Y_S . \Phi_T = \Phi_S . X_{cs} . \Phi_T + No', \tag{3.8}$$

[7] Note that the use of this kind of sparse measurement matrices refers to the analog CS, called also Low-rate CS [17].

[8] According to the analysis made by [27], the coherence between the sensing/compression matrix and the sparsifying one is determined by the column vectors of Φ. Since Ψ is an orthonormal basis matrix, we require that $\Phi^{tr}\Phi \approx I$ in order to get a small coherence value, i.e. the column vectors of Φ are required to be approximately orthogonal.

where $No' \in \mathbb{R}^{M_s \times M_t}$ and $No' = No.\Phi_T$.

In this equation, Φ_T has a sparseness structure similar to that of Φ_S since it holds a single '1' in each column and at most a single '1' in each row. This multiplication consists of randomly selecting M_t moments among Tcs moments for which a given active sensor will transmit its readings to the sink. To summarize, if it is selected as a transmitting source node among the Ms selected nodes, this sensor node has just to collect measurements according to a designated temporal schedule and transmit them to the sink. As a result, we obtain a much lower number of measurements in $Y \in \mathbb{R}^{M_s \times M_t}$ compared to the original $2D$ signal $X_{cs} \in \mathbb{R}^{N \times T_{cs}}$.

3.3.2 Kronecker sparsifying basis

To reach an accurate recovery of the received $2D$ compressed signal Y, we take advantage of the space-time correlation existing in the original signal X_{cs} to highlight its sparseness in its two dimensions. In fact, each of the signal dimensions owns a sparse representation in a proper transform domain, denoted as $\Psi_S \in \mathbb{R}^{N \times N}$ for the spatial basis and $\Psi_T \in \mathbb{R}^{T_{cs} \times T_{cs}}$ for the temporal one. Thus, we have:

$$X_{cs} = \Psi_S.\alpha.\Psi_T, \tag{3.9}$$

where $\alpha \in \mathbb{R}^{N \times T_{cs}}$ is the $2D$-sparse representation of X_{cs} in Ψ_S and Ψ_T. The determination of these two bases is very important since they are deeply involved in the reconstruction step of CS as shown in expressions (2.4) and (2.7). Therefore, we detail hereafter how these transformation bases have been implemented.

3.3.2.1 Signal transformation in the spatial domain

To construct the spatial sparsifying basis Ψ_S, we resort to the online estimation PCA approach merged with the covariogram theory, proposed in [21]. In order to estimate the spatial correlation inherent in the signal X_{cs}, this related approach relies on the computation of the experimental variogram $\gamma_{exp}(d)$ using the learning data matrix

X_{lp}. That is, for a given geographical distance $d > 0$, we have:

$$\gamma_{exp}(d) = \frac{1}{2N_d} \sum_{p' \ s.t. \ \|p-p'\|_2=d} [x_{lp}(p) - x_{lp}(p')]^2, \tag{3.10}$$

where $x_{lp}(p)$ denotes the data sample of X_{lp} sensed from the space location p, and N_d represents the number of pairs (p, p') that are separated by the same distance d. According to [79], for a stationary data field, the experimental variogram can be computed through the experimental covariance variables of the signal of interest. That is, for a given geographical distance d, we have:

$$\gamma_{exp}(d) = \frac{C_{exp}(0)}{N_0} - \frac{1}{N_d} \sum_{p' \ s.t. \ \|p-p'\|_2=d} C_{exp}(d), \tag{3.11}$$

where $C_{exp}(0)/N_0$ denotes the average variance computed from the considered data samples $(x_{lp})_p$ since $C_{exp}(0)$ denotes the sum of the covariances that correspond to zero distances, N_0 denotes the number of the considered data samples $(x_{lp})_p$, and $C_{exp}(d)$ denotes the covariance computed from the pair (p, p').

To estimate the average variogram $\gamma_{exp}(d)$ for a set of samples pairs on an irregular grid with distance d, Jindal *et al.*, in [80, Section. 2], have stated a simple and detailed method for that:

1- For each pair of samples, distance d between them and the squared difference between their data values $[x_{lp}(p) - x_{lp}(p')]^2$ are calculated.

2- The entire range of distances is divided into discrete contiguous intervals[9].

3- Attribute each of the pair of samples (p, p') to one of the distance intervals, then calculate the average variogram for each interval through the division of the sum of the squared-differences between data values by the total number of pairs lying in that distance interval.

Generally, performing (3.10) or (3.11) is followed by the search of the theoretical variogram values $\gamma_{th}(d)$, which represent the values of the variogram expression model that is chosen from a set of predefined variogram models (such as the spherical, gaussian, circular, etc.) so as to fit the best the experimental values $\gamma_{exp}(d)$. This step is done in order to provide correlation information between locations where there is no gathered data. Indeed, the expression (3.10) or (3.11) estimates the average of the

[9] The interval size can be fixed according to the average distance to the nearest neighbor.

experimental variograms for distances d using only the available data samples.
Once the best suited variogram fit γ_{th} is obtained, it is integrated in the computation of the covariogram matrix $\Sigma_c \in \mathbb{R}^{N \times N}$ using the following expression, where the element (i, j) of Σ_c can be written as:

$$\sigma_{c\ i,j} = sill - \gamma_{th}(d_{i,j}). \tag{3.12}$$

In this equation, $sill = \lim_{d \to \infty} \gamma_{th}(d)$ is a parameter that is obtained during the selection process of the suitable variogram model and $d_{i,j}$ is the geographical distance between sensor i and sensor j.
Once Σ_c is estimated, we consider the orthonormal basis Ψ_S whose columns are the eigenvectors of Σ_c, corresponding to the eigenvalues sorted in decreasing order. As it will be validated with simulations in the next section, this combined covariogram-PCA method exploits well the correlation among sensors[10] and makes improvements compared to the sample covariance-PCA method used in [51]. The advantage of this method is that Ψ_S is dynamically adapted to the signal model and is not fixed for all the sensing periods.

3.3.2.2 Signal transformation in the temporal domain

Regarding the temporal basis Ψ_T, we use the Discrete Cosine Transform (DCT), given by the following expression [20]:

$$\Psi_{T_{i,j}} = C_{dct}.\cos((i-1)(1+2(j-1))\frac{\pi}{2T_{cs}}), \tag{3.13}$$

where $C_{dct} = \sqrt{1/T_{cs}}$ if $i = 1$ and $\sqrt{2/T_{cs}}$ otherwise. The DCT is very similar to the Discrete Fourier Transform (DFT) in the sense that it gives a spectral analysis of the data [70]. The DCT makes a sparse signal by concentrating most of its information into few low frequency components. The remaining high frequency components tend to be weak values and become less important, and thus they can be removed without visual losses [20]. Both of the bases DCT and DFT worked well with our approach and gave similar results.

It is noteworthy that in contrast to [21], where new Ψ_S and Ψ_T are computed in each

[10] As it has been stated in [21], the proposed method works well with signals that are non-stationary.

time slot t to recover the correspondent data vector (column vector t of X_{cs})[11], in this work, these bases are calculated once to rebuild the entire sensing data X_{cs}.

Note that the data learning X_{lp} is used only for the first sensing period T_{cs}^{1}, where the sink node does not have information corresponding to the sensor nodes. Yet, for the following sensing periods T_{cs}^{T}, i.e. $T > 1$, it makes use of the just previous recovered data matrix $\widehat{X}_{cs}^{T-1}$ of the previous sensing period T_{cs}^{T-1} to adaptively estimate both the compression matrix and the transformation basis that will be used during the current sensing period T_{cs}^{T}. Besides, these appropriate matrices are computed, known and used only by the sink. As we can see, our algorithm imposes neither inter-sensor communication nor on-sensor computation. Hence, our STCS algorithm is characterized by its simple encoding and complex decoding as required in the CS for WSNs. Figure 3.3 illustrates a flowchart that simplifies the design of the proposed approach. For simplicity reasons, the index referring to the ordering of the sensing periods is used only in this part to better explain the working of the proposed approach.

3.3.3 From matricial product to kronecker product

As illustrated in expressions (2.6) and (2.7), the resolution of standard CS is formulated with x and y in vector form. Therefore, we use tools from linear algebra in order to reformulate the $2D$ problem as a $1D$ problem. It is worth noting that this conversion does not lose or change any information and preserves the intra and inter-correlations [36]. Using [36, Eq. 13 and Eq. 14], we consider the vec(.) function, which converts a $P \times Q$ matrix to a $P.Q$ vector by vectorizing it by column. Then, we can write : $vec(Xcs) = (x_{cs}(1,1), ..., x_{cs}(N,1), x_{cs}(1,2), ..., x_{cs}(N,2), ..., x_{cs}(1,Tcs), ..., x_{cs}(N,Tcs))^{tr}$, and (3.8) becomes:

$$\begin{aligned} y &= (\Phi_T^{tr} \otimes \Phi_S).vec(X_{cs}) + vec(No') \\ &= \Phi.vec(X_{cs}) + vec(No'), \end{aligned} \tag{3.14}$$

where $\Phi \in \mathbb{R}^{M_t.M_s \times T_{cs}.N}$ is the kronecker product between Φ_S and the transpose of Φ_T and $y \in \mathbb{R}^{M_t.M_s \times 1}$. As in [28], we can obtain a single sparsifying basis Ψ for an

[11] In [34], authors have used a kind of sliding window processing that covers the data of $W < T_{cs}$ successive time slots to estimate the data vector of the current time slot and re-estimate those of the previous $W - 1$ time slots.

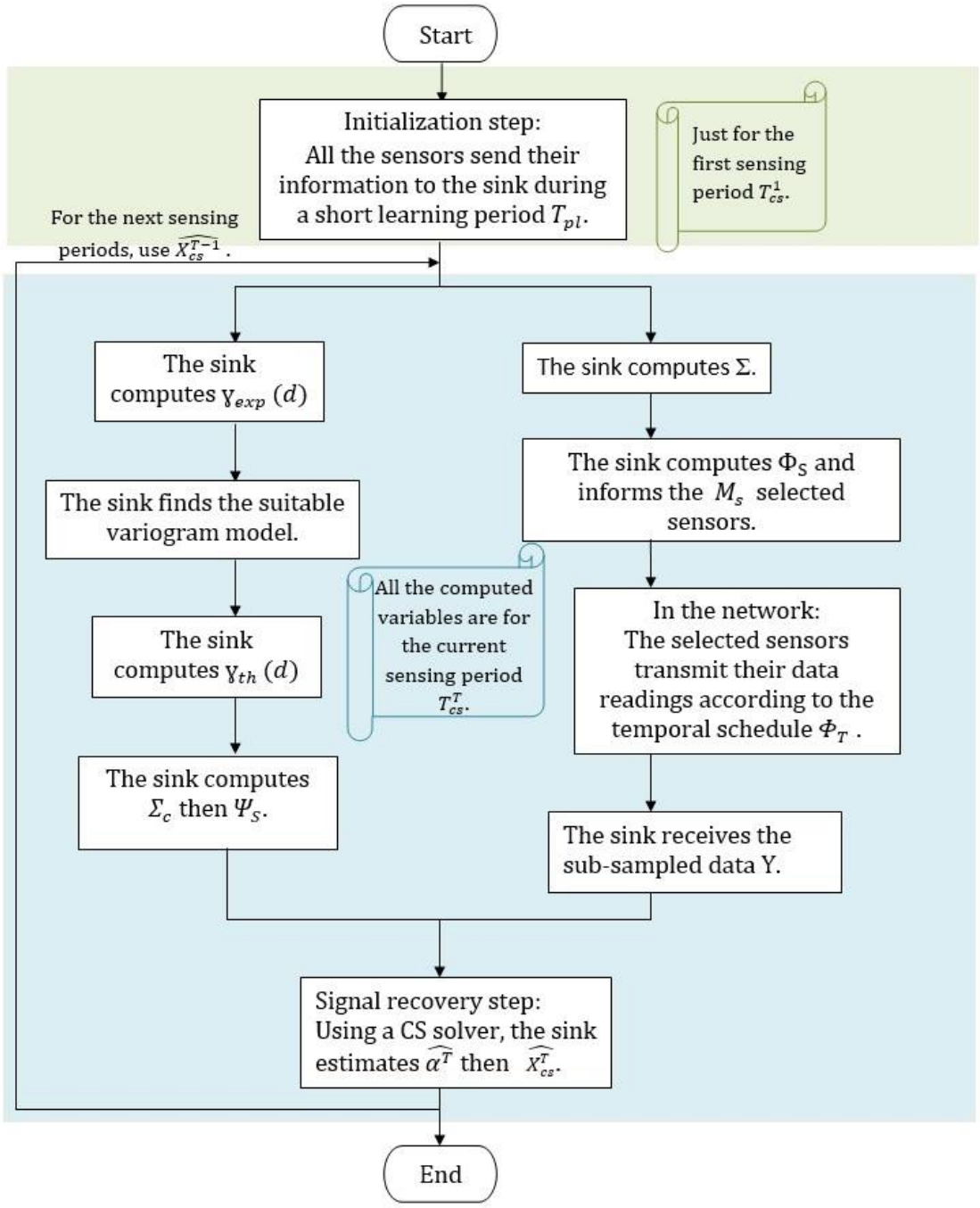


Fig. 3.3. A flowchart simplifying the design of the proposed approach.

entire 2-dimensional signal as the Kronecker product of sparsifying bases for each of its 2-sections. Considering the same tools used previously in (3.14) with the projection matrices, we carry out the same conversion to (3.9):

$$\begin{aligned} vec(X_{cs}) &= (\Psi_T^{tr} \otimes \Psi_S).vec(\alpha) \\ &= \Psi.vec(\alpha), \end{aligned} \tag{3.15}$$

where $vec(\alpha)$ is a vector-reshaped representation of α and $\Psi \in \mathbb{R}^{T_{cs}.N \times T_{cs}.N}$ is the joint sparsifying basis over space and time. The expressions (3.14) and (3.15) take us back to the standard CS formulation:

$$y = \Phi.\Psi.vec(\alpha) + vec(No'). \tag{3.16}$$

3.4 Numerical Results

In this section, we analyze the performance of STCS and we compare our results to those of CS²-collector and CB-CS. Then, we evaluate the STCS-RA with respect to different parameter values of β, see equation (3.4). The metrics, that we use for the simulations, are the normalized Mean Squared Error (MSE) and the Compression Ratio (η) defined as follows:

$$MSE = \frac{\|X_{cs} - \widehat{X}_{cs}\|_F^2}{\|X_{cs}\|_F^2} \quad \text{and} \quad \eta = \frac{(\upsilon - \delta)}{\upsilon}, \tag{3.17}$$

as well as E, the average consumed energy per sensor per time slot (μJ) [21]. X_{cs} and $\widehat{X}_{cs}$ represent respectively the sensed $2D$ signal before compression and the $2D$ recovered one by the sink for a given sensing period, whereas, $\|.\|_F$ is the Frobenius norm. υ and δ present respectively the number of elements in X_{cs} and in the $2D$ compressed data Y.

For the network parameters, we consider $N = 50$, $T_{cs} = 90$ and the observation area size is 100×100 units. Regarding η, we vary M_t between 9 and T_{cs}, and M_s between 5 and N.

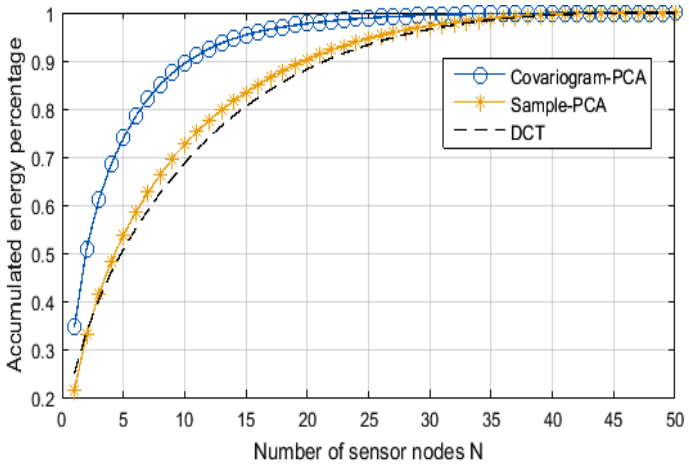


Fig. 3.4. The signal accumulated energy percentage with different sparsifying basis for ($\rho = 0.9$, $\gamma = 2$).

To begin, before going into the CS-based approaches performance comparison, the signal accumulated energy percentage with the studied transformation bases are calculated, according to the method of [57], then depicted in Figures 3.4 and 3.5, with the variation of the spatial correlation parameter γ. As we can note, the combined covariogram-PCA transformation basis sparsifies better the signal than the sample-PCA transformation basis for both cases. As an example, when $\gamma = 5$, while approximately 10% of the covariogram-PCA coefficients assemble 80% of the signal energy, approximately 10% of the sample-PCA coefficients assemble less than 70% of the signal energy. Here, we added the curve for the DCT basis to be a reference, since the DCT matrix is considered as a standard transformation basis in the CS theory.

In the next simulations, we compare our algorithm STCS with the CS²-collector in terms of normalized MSE and η with the variation of the spatial correlation γ parameter. From Figure 3.6, we can see that the reconstruction accuracy (lower MSE) increases with the number of measurements M_t and Ms (lower η), and STCS provides considerable improvements compared to CS²-collector across the entire range of η, especially when the transmitted signal is correlated in space. This is due to the fact that with STCS we exploit well the spatial dependency in the signal thanks to the node selection strategy and the covariogram-PCA method of [21] to construct Φ_S and Ψ_S. This is different to CS²-collector that chooses to select nodes randomly and uses a simple DCT matrix as Ψ_S. Even for the moderately correlated signal in space,

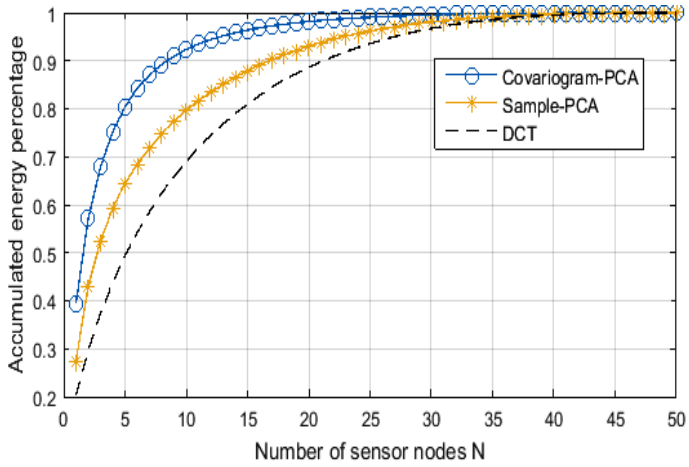


Fig. 3.5. The signal accumulated energy percentage with different sparsifying basis for ($\rho = 0.9$, $\gamma = 5$).

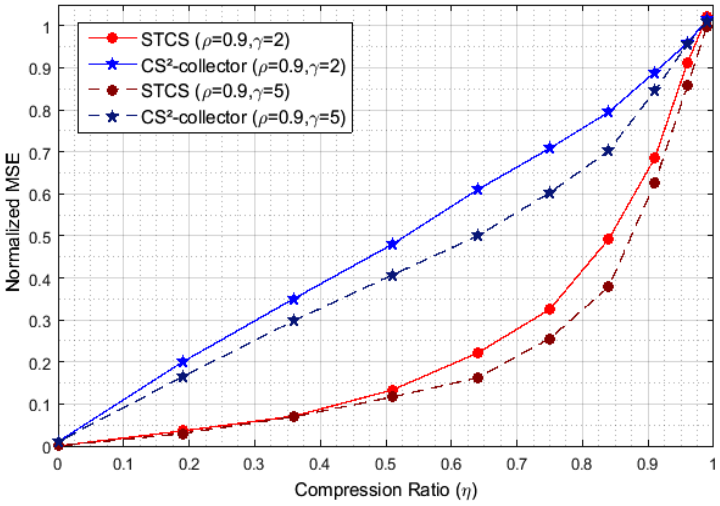


Fig. 3.6. A performance comparison in terms of reconstruction error between STCS and CS²-collector for ($\rho = 0.9$, $\gamma = 2$) and ($\rho = 0.9$, $\gamma = 5$).

STCS outperforms CS²-collector, which linearly combines all the T_{cs} sensors readings along with a dense random Gaussian matrix. Different from the CS²-collector, our STCS sub-samples the sensors readings by taking only a fraction of them.

Regarding E, in order to be comparable with [21], we use the same hardware implementation, i.e, an MSP430 Micro-Controller with CC2420 radio. In Figure 3.7, we compare the average consumed energy per sensor per time slot of our proposed STCS

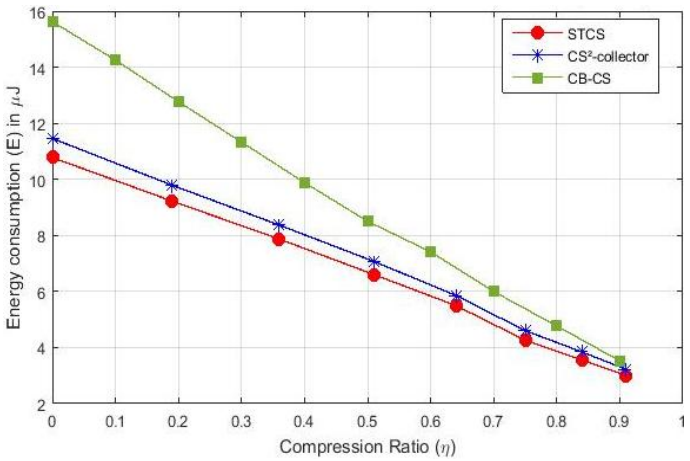


Fig. 3.7. A performance comparison in terms of energy consumption between STCS, CS²-collector and CB-CS.

with CB-CS and CS²-collector. The energy consumption in Figure 3.7 takes into account the energy for transmission, reception and on sensor processing. A noticeable observation in Figure 3.7 is that our STCS scheme consumes much less energy than the other schemes especially the CB-CS due to the communication cost. Regarding the computation cost, the reason that makes STCS reduce its total energy consumption compared to CS²-collector is that the STCS has no on-board computation thanks to the sparse combination of their compression matrices compared to the CS²-collector (it requires $M_t \times [T_{cs}$ multiplications$+(T_{cs}-1)$ additions$]$ in each sensor since the temporal compression matrix Φ_T is a dense Gaussian matrix and the temporal compression precedes the spatial one). We consider respectively 395 and 184 clock cycles for the multiplication and the addition operations [81], and E_{cc} =0.726 nJ the energy consumption per clock cycle for an MSP430F1612 [21]. The minimization or even the cancellation of the number of operations improves the runtime and consequently optimizes the response in real time.

Obviously, the radio consumes the bulk of the total power consumption of WSN systems as shown when comparing with CB-CS but considering as well the computation cost can be more beneficial for further improving the overall energy efficiency as shown when comparing with CS²-collector.

Figures 3.8 and 3.9 depict the trade-off between the normalized MSE and E for different values of β (note that $\beta = 0$ corresponds to STCS according to (3.2)). In order to

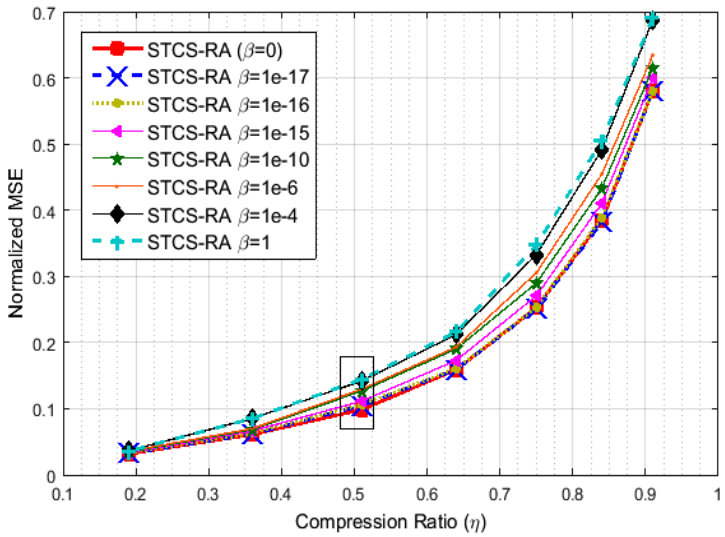


Fig. 3.8. Normalized MSE for STCS ($\beta = 0$) and STCS-RA with respect to different β for ($\rho = 0.9$, $\gamma = 5$).

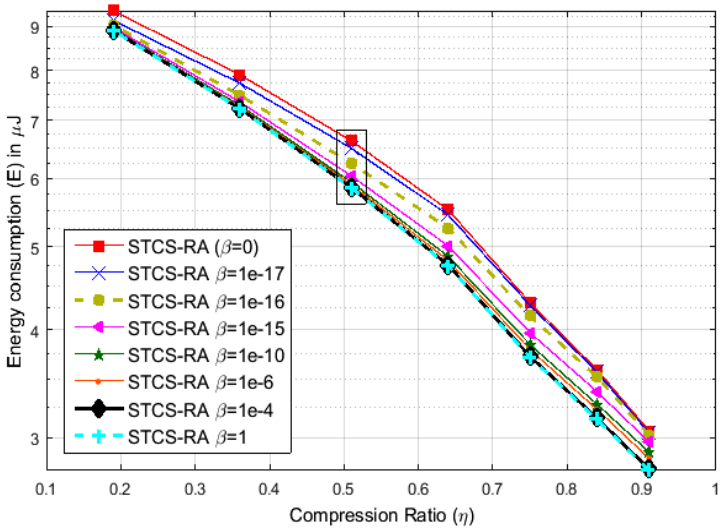


Fig. 3.9. Energy consumption E for STCS ($\beta = 0$) and STCS-RA with respect to different β for ($\rho = 0.9$, $\gamma = 5$).

vary the normalized MSE, we changed η from 0.19 to 0.91. For each case, the energy is calculated and then depicted in Figure 3.9. We note the improvement of sensor lifetime by considering the routing in the metric. For example, when $M_s = 35$ ($\eta = 0.51$), the normalized MSE is about 0.11. Thus, we can save 11,11% of energy when

$\beta = 10^{-15}$. As it can be seen, the curves for $\beta = 10^{-4}$ and $\beta = 1$ are superposed in Figure 3.9 but slightly different in Figure 3.8. It means that the performance is very dependent to the number of hops, and because it is an integer value, among all the paths with the same hop number, the one giving the best correlation properties is selected. Another observation from these two plots is to say that the optimum sensor selection is sensitive to the correlation criteria (γ). Therefore, for small M_s (big Compression Ratio), it is important to weaken the effect of routing by reducing β.
As a perspective, it is possible to include the residual energy of sensors in the cost function (3.4). In this way, even though the energy consumption will not be minimized, the overall lifetime can be extended.

3.5 Conclusion

Motivated by reducing efficiently the number of representative measurements to be transmitted to the sink node, thanks to the redundancy nature of most WSNs signals, we addressed in this chapter the STCS approach. We proposed a joint space-time compression scheme that adaptively learns the signal model from the past received data to schedule when and where to sample the $2D$ time-varying spatial field. Then, we recover the entire $2D$ signal from the small number of measurements using appropriate transformation bases, that can well sparsify the signal according to the correlation structure inherent on it.
Characterized by a much lower number of transmissions and no on-sensor computation, STCS reduces the energy consumption compared to other CS-based schemes, while still achieving appealing reconstruction performance. This trade-off between energy saving and reconstruction accuracy has been further improved with the STCS-RA, which takes into account the routing in the representative node selection process.

Chapter 4

Robust Data Recovery in Wireless Sensor Network: A Learning-Based Matrix Completion Framework

Contents

4.1 Introduction

In some applications, especially the densely deployed WSNs, the sensed data is in general highly correlated, and redundancy exists between sensor nodes belonging to the same geographic area. To enhance the network management, nodes can be arranged into groups or clusters. Since they are monitoring the same targets or events, collecting raw data from all cluster members becomes inefficient and energy wasteful. Therefore, as a sequel of the previous chapter 3, a sufficient subset of nodes can be selected from each group to be the representative of the whole network. These active nodes deliver their readings to the sink under a compression ratio, while the rest of nodes remain silent and do not participate in the data sensing operation.

The CS is an interesting proposal since it reduces the number of active agents at a given time slot, while remaining able to recover the sensing data. However, to reach a sufficiently satisfying data interpolation quality with a higher compression rate, i.e. fewer delivered data readings, the signal correlation should be fully captured simultaneously in both space and time dimensions. To do so with the kronecker CS framework, as we have already seen in the previous chapter 3, tools from linear algebra are still be needed in order to reformulate the data matrix into the vector form since the standard resolution of CS is still formulated in the $1D$ form. Without the need of computing an adaptive sparsifying basis Ψ, the MC, viewed as an extension of the CS, has emerged recently using another type of structural sparsity[1] [44], which is the matrix low rank property [23]. Since it treats the data matrix as a genuine matrix, MC can take advantage of the correlation in its two dimensions and capture more information[2].

In this chapter, we carry on with the twofold data compression scenario, where we firstly assume that part of nodes do not sense the environment at all. We can consider that these sensors are inactive or idle for a long period or that these nodes are absent. The second compression level is that, at each time slot, only a subset of the active nodes, referred to as the transmitting source ones, send their sensing data to the sink. Note that different from the previous work, where according to the temporal

[1] A low-rank matrix holds singular values composing a sparse spectrum.

[2] In [82, Fig. 3], we have illustrated that a simple MC-based approach requires a smaller fraction of sensor node readings to reach the same data recovery accuracy.

sampling pattern (3.8) there are several time slots during which no data is transmitted, in this work, at every time slot, we ensure the transmission of a number of data readings sensed from different locations belonging to different clusters of the monitored network area. This kind of strategies not only minimizes the energy cost and extend the network lifetime, but also helps to avoid other problems such as the traffic congestion [50,83].

Yet, the application of these atypical high-loss scenarios leads to a significant number of empty rows in the received data matrix[3], which completely disagrees with MC fundamentals. In fact, since MC approaches are based on the minimization of the matrix rank, they become useless when there is any empty row or empty column in the matrix. Indeed, MC techniques have been conceived to recover matrix containing random missing elements [84]. Even though the existence of the inactive sensor nodes has already been considered, in the previous work of chapter 3, the recovery of their missing data has been achieved using the CS technique with the Kronecher framework.

In the state-of-art of MC-based algorithms in WSNs, to the best of our knowledge, [24] is the only paper who dealt with the case where there are some missing rows in the received data matrix. They appeal a spatial pre-interpolation technique that recovers data from neighboring sensor nodes. However, as the number of active nodes decreases, we face absent nodes having absent neighbor sensors as well. Thus, this framework becomes unable to recover the data rows of these *isolated* sensor nodes. Although this approach is interesting, it seems to be not well suited for the addressed scenario and fails to take into account the existence of the *isolated* sensor nodes (absent nodes having all their neighbors absent). In this context, we present our developed scheme, which firstly, schedules the sampling pattern after efficiently identifying the different clusters and their representative nodes. Secondly, it treats the case of high compression ratios with a considerable number of inactive sensor nodes (empty rows) using a combination of three different interpolation techniques.
The main contributions of this chapter are summarized as follows:

- We generate a synthetic space-time signal composed of different Gaussians, each of which presents a cluster of wireless nodes. As in all the WSNs signals' profiles, the portions are correlated in space and time, where spatial and temporal

[3] A row (resp. column) is called an empty row (resp. column) if and only if all the values of the row (resp. column) are un-sampled.

correlation parameters differ from one Gaussian (portion) to another and can be separately adjusted.

- To perform an adaptive data gathering, a preliminary phase is established, where nodes are arranged into a number of clusters. Then, in order to equitably involve all the detected clusters in the sensing schedule and ensure the diversity in the transmitted data, in each time slot, using the same percentage and according to a given sampling ratio, a subset of nodes is picked from each cluster to ensure data sensing.
- For the reconstruction part, we propose to use three different techniques to accurately rebuild the entire data matrix. In the first step, we fill the missing readings of the active sensor nodes by applying the MC. Then, we carry on with the spatial pre-interpolation to handle a part of the empty rows while adjusting the topology matrix to the presence of the disjoint clusters in the monitored field. Finally, we recover the rows of the *isolated* sensor nodes using a minimization-based interpolation technique with a spatial correlation matrix.
- Through extensive simulations, we show that the proposed framework outperforms other existing techniques in the literature, especially when the number of inactive nodes increases.

The remainder of the chapter is organized as follows. The next section discusses the problem formulation of this work. In section 4.3, we present the signal model that we used for the evaluation of our approach. Then, in section 4.4, firstly, we introduce the efficient clustering method that we propose. Secondly, we describe in detail our strategy for an adaptive data sampling. Section 4.5 is dedicated to the data reconstruction framework. Before concluding the chapter in section 4.7, we carry out, in section 4.6, with extensive simulations in order to evaluate the performance of the proposed approach.

4.2 Problem Formulation

Consider a WSN composed of a set $\mathcal{N}_\int = \{1, \ldots, N\}$ of N sensor nodes. Let $X \in \mathbb{R}^{N \times T}$ denote the data matrix that contains measurements collected by the set $\mathcal{N}_\int$

during a sensing period of length T time slots. Precisely, the entry in the i^{th} row and t^{th} column of X, $x_{i,t}$, represents the t^{th} data reading ($t \in [1,T]$) sensed by the i^{th} node ($i \in \mathcal{N}_\int$). The considered scenario aims to obtain all sensor nodes readings, X, through the use of a small subset $\mathcal{N}_{rep} = \{1, \ldots, N_{rep} \ll N\}$ of active sensors, denoted by representative sensor nodes. It is worth mentioning that the number of active sensors is relatively small compared to the number of inactive ones. Specifically, decreasing the number of active sensors can likely generate a set of absent sensors that have also all their neighbors absent as well. We call them *isolated* (IS) sensor nodes.

We propose to group together sensor nodes having similar readings in the same cluster using a spectral clustering technique. In fact, the whole network is organized as follows: $\mathcal{N}_\int = \bigcup_{j=1}^{J} CL_j$ and $N = \sum_{j=1}^{J} cl_j$, where cl_j is the number of sensor nodes belonging to CL_j ($cl_j = card(CL_j)$), J is the number of detected clusters and CL_j is the cluster j. It will be shown, in the sequel, that the representative node selection as well as the data transmission schedule depend on the detected clusters.

To further reduce energy consumption, the representative sensors do not transmit their raw data to the sink. Instead, they trade on the data sensing along the T time slots and deliver a part of their readings according to a given compression ratio, that is, $m < N_{rep}$ readings rather than N_{rep} readings per time slot. Consequently, the received data matrix $M \in \mathbb{R}^{N \times T}$ is composed of N_{rep} partially empty data rows and $(N - N_{rep})$ completely empty data rows. Note that to replace any missing entry in M, we set a "zero" as a placeholder. We use a binary sample matrix $\Omega_M \in \mathbb{R}^{N \times T}$ that we call sensing and transmitting schedule to indicate, in each time slots t, which nodes sense and transmit measurements. That is,

$$\Omega_{M(i,t)} = \begin{cases} 1 & \text{if } x_{i,t} \text{ is available} \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

Note that $x_{i,t}$ is available, when the location i is sampled and transmitted at time slot t. We refer to a location by i when it is sampled by the sensor node i. Hence, the incomplete delivered data matrix M can be expressed as follows:

$$M = X. * \Omega_M, \tag{4.2}$$

where $\cdot *$ represents a Hadamard product of two matrices.

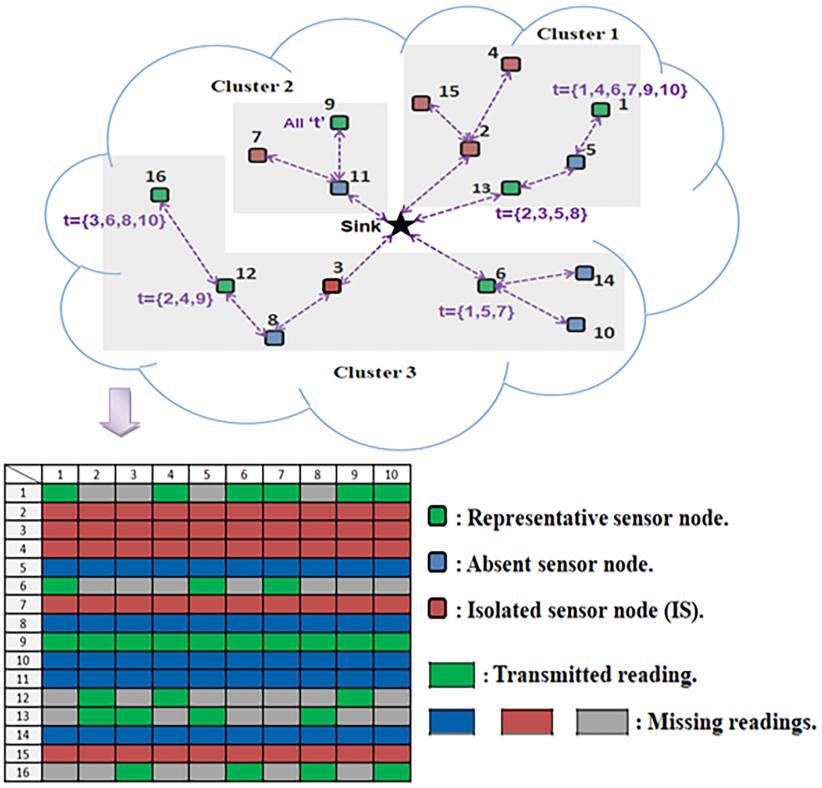


Fig. 4.1. An illustrative miniature WSN with the resulting transmitted data matrix M.

The first aim of our work is to well identify the matrix Ω_M as it represents the sampling schedule, which is of prime importance in the recovery performance.

The second aim of our work is to successfully recover all the missing entries using a limited number of received readings. Therefore, we opted for the MC technique because of its numerous benefits. Indeed, the application of MC with the existence of a significant number of empty rows is still a challenging task to tackle since the presence of empty rows or columns impedes the MC reconstruction. Thereby, we propose in this chapter a novel interpolation technique that will be annexed to the MC one in order to recover the empty rows. It is noteworthy that the MC, as the first step in the reconstruction operation, is an important part since the performance of the subsequent proposed interpolation technique depends on the recovery accuracy of the MC. Figure 4.1 illustrates an example of a WSN consisting of $N = 16$ sensor

nodes, among which $N_{rep} = 6$ sensor nodes are selected to be active. The proposed combined reconstruction approach targets to fill all the missing entries corresponding to the non-transmitted readings.

4.3 Multi-Gaussian Signal Model

4.3.1 The Signal Generation

In this section, we investigate the generation of a synthetic signal composed of different Gaussians, each of which presents a portion of the whole monitored geographic area. Each portion of the signal is correlated in space and time, where the spatial correlation as well as the temporal correlation parameters differ from one Gaussian to another.

The proposed signal model is inspired by [77] that has introduced the solution of reproducing a signal retaining the behavior of a given real world data by adjusting the correlations parameters. In their model, all the generated samples of the whole signal are Gaussian random variables with a zero mean and a variance following the spatial correlation function used in the signal generation. Indeed, according to [77, Eq. 14], for $p = (x, y)$ with $x = \{1, 2, ..., N_D\}$ and $y = \{1, 2, ..., M_D\}$, representing a space point of a sensor grid of $N_D \times M_D$ points, the resulting variance of $z(p, t)$ following algorithm 1 is $\sigma^2_{z(p,t)} = \sum_{i=1}^{N_D} \sum_{j=1}^{M_D} rs(x - i, y - j)^2$. However, in this chapter, we consider heterogeneous fields that are divided into a number of regions. Each one is modeled by a specific Gaussian (mean, variance) and different correlation characteristic. The number of different Gaussians as well as their distribution in the field can be fixed or defined according to the kind of the signal one wants to reproduce. Thereupon, this method represents an effective alternative to the real world signals.

As in chapter 3, to generate the signal of interest, we suppose that $D = [-x_D, x_D] \times [-y_D, y_D]$ is the space domain, where x and y are the space coordinates. Then, we consider that we have H different regions, where D_h is the space domain of region $h = 1, 2, \ldots, H$, and $D = \bigcup_{h=1}^{H} D_h$. Without loss of generality, for a given pair (ρ_h, γ_h) of specific temporal and spatial correlation parameter values, we suppose that algorithm 1 of chapter 3 describes how to generate a correlated portion of the signal $z_h(p_h, t) : D_h \times \mathcal{T} \rightarrow \mathbb{R}$ representing one region, where $\mathcal{T}$ is the time domain and

p_h is a point in (x, y) plane corresponding to region h. Note that the signal of the whole area is the combination of all the generated portions. The resulting z_h holds generated samples with zero mean and variance that depends to the performed spatial correlation function. Accordingly, in order to obtain an heterogeneous signal field for the entire network, for each region h, we enforce a non-zero mean $\eta_h \in \mathbb{R}_{\neq 0}$ to its corresponding generated signal z_h as follows:

$$z_h^{'}(p_h, t) = z_h(p_h, t) + \eta_h. \tag{4.3}$$

In addition to the mean, the variance amplitude σ_h^2 of the signal field that one wants to produce can be tuned by multiplying the samples $z_h(p_h, t)$ by a constant parameter $cst_{\sigma_h} > 1$, according to the following expression:

$$z_h^{'}(p_h, t) = cst_{\sigma_h}.z_h(p_h, t) + \eta_h. \tag{4.4}$$

Algorithm 1 followed by (4.3) or (4.4) outlines how to produce a portion $z_h^{'}(p_h, t) : D_h \times \mathcal{T} \rightarrow \mathbb{R}$ of the whole signal field $z^{'}(p, t) : D \times \mathcal{T} \rightarrow \mathbb{R}$, which represents the (x, y) signal. Similarly to what we have done in section 3.2.2, $z^{'}(p, t)$ represents a $3D$ matrix of size $(2y_D \times 2x_D \times T)$, and the data matrix of interest, X, denotes the $2D$ signal discretized from $z^{'}$ by the N sensor nodes along the T time slots.

Figure 4.2 illustrates an example of an area of size $100m \times 100m$ monitored by $N = 50$ sensor nodes. We can notice through the colors that this field is divided into three different regions $(H = 3)$ presented by three different Gaussians.

4.3.2 The Low-Rank feature

To ensure the use of the MC, the manipulated data matrix should exhibit a low rank or approximately low-rank structure. To do so, one can use the SVD method [55]. In fact, any real $N \times T$ matrix X can be written as follows:

$$X = U \Lambda V^T, \tag{4.5}$$

where $V \in \mathbb{R}^{T \times T}$ and $U \in \mathbb{R}^{N \times N}$ are two unitary matrices and $\Lambda \in \mathbb{R}^{N \times T}$ is a diagonal matrix assembling the singular values τ_i of X. Typically, $\tau_1, \tau_2, ..., \tau_r$ are

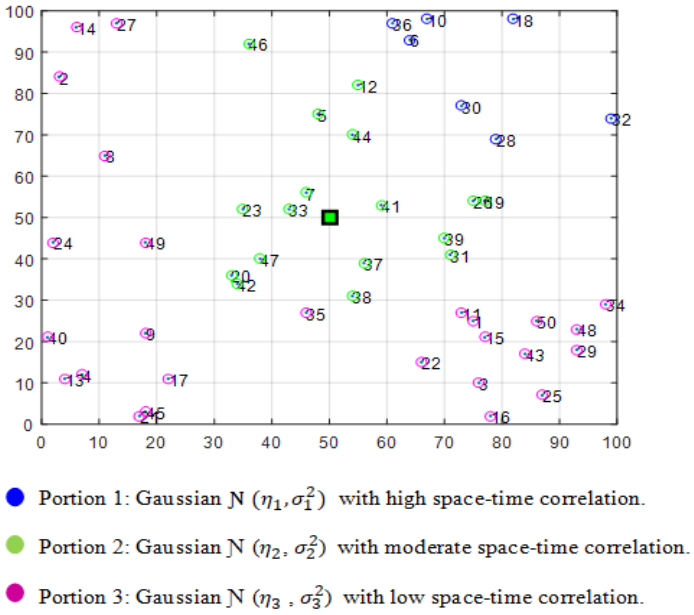


Fig. 4.2. An example of a monitored area composed of three portions, each of which is presented by a different Gaussian.

arranged in a decreasing order so that $\tau_i \geq \tau_{i+1}$, where r denotes the rank of X. If we find out that the top l singular values of the data matrix X occupy the near total or the total energy, then X holds the low rank feature. The metric that we use to check this property is the fraction of the nuclear norm captured by the top l singular values [55]:

$$g(l) = \frac{\sum_{i=1}^{l} \tau_i}{\| X \|_*} = \frac{\sum_{i=1}^{l} \tau_i}{\sum_{i=1}^{r} \tau_i}. \tag{4.6}$$

As we have mentioned before, the low rank property, inherent in the signal, enables the use of the MC tools to recover the raw data matrix from the received entries. Figure 4.3 plots the fraction of the total variance captured by the top l singular values for a signal generated from the monitored field presented in Figure 4.2. The signal generation parameters for this example are summarized in Table. 4.1. We note from the plot that the top 5 singular values capture nearly 93% of the nuclear norm, which indicates that the signal matrix X has a very good low-rank approximation. Hence, we are able to apply the MC technique.

Table 4.1: Signal generation Parameters

Parameter	Portion 1	Portion 2	Portion 3
η_h	35	20	5
ρ_h	0.9	0.7	0.5
γ_h	7	5	2

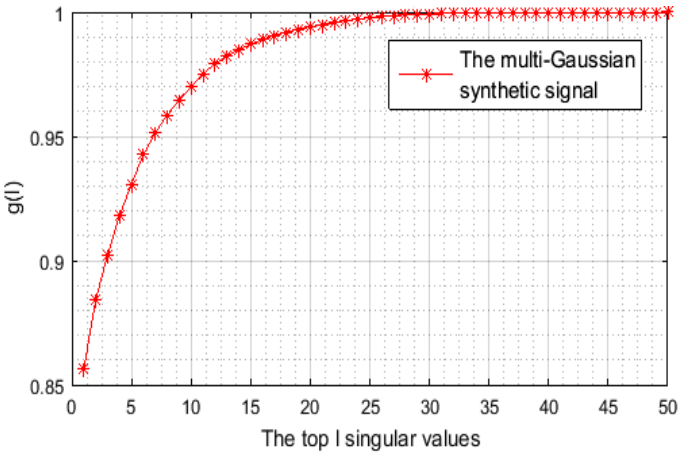


Fig. 4.3. Fraction captured by the top l singular values for a multi-Gaussian synthetic signal, generated using the values of Table. 4.1.

Even though the low-rank feature studied above may indicate the existence of redundancy and dependency structure in the data matrix X, which reflects the spatial correlation and the temporal correlation properties of the data, we have provided as well, in the appendix B, a separate study for each of them; B.1 and B.2.

4.4 Sampling Pattern

4.4.1 Clusters Detection

In this part, we investigate the partition of the deployed sensor nodes into J clusters. The main reason for partitioning the nodes is to involve all the detected clusters in the data sensing. In the conventional MC, it is well-known that transmitting source nodes are selected in a purely random way during the T time slots. This kind of selection can disregard sensors belonging to small clusters, which can heavily deteriorate the recovery process. However, if we ensure that all the clusters contribute in the data sensing and transmission process, we can fortify the diversity in the delivered data

set and thus enhance the data reconstruction quality. Therefore, for each time slot t, according to a given compression ratio and using the same shared percentage, a set of sensor nodes is picked from each cluster to form the sampling and transmission schedule. It will be shown, in the simulation section, that taking into account the detected clusters during the sampling process significantly enhances the data recovery performance, especially for high compression ratios. Indeed, our aim is to partition the sensor nodes into different clusters, where nodes in the same cluster have similar readings. Namely, we attempt to minimize the inter-clusters similarity and maximize the intra-clusters similarity, and such a successful grouping can be achieved using the normalized spectral clustering, unlike the unnormalized one that implements only the first objective [85].

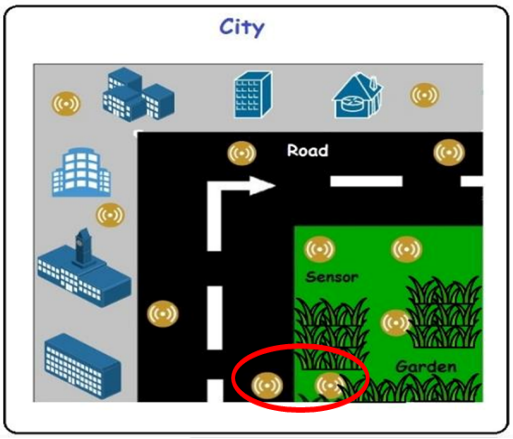


Fig. 4.4. Civilian and habitation deployment areas for sensor nodes.

Usually, sensor nodes, which are situated spatially close to each other, have similar readings. Nevertheless, there are some cases, where nearby nodes are separated by a certain barrier and have readings relatively different from each other. In the example of Figure 4.4, sensor nodes are deployed in a city to monitor the air pollution. Suppose that we have a public garden located next to a road. Hence, the nearby nodes, which are placed on the two different sides of the borders, do not necessarily have similar readings. Therefore, to cluster the nodes, the sink relies on their delivered readings[4] and considers the set of data vectors, $\chi_{lp} = \{x^{tr}_{lp\,1}, x^{tr}_{lp\,2}, \ldots, x^{tr}_{lp\,N}\}$, that we want to partition into J clusters. The spectral clustering technique performs data clustering

[4] As in chapter 3 with the STCS, at the initialization, we let all the sensor nodes send their information during a short learning period $T_{lp} \ll T$.

and treats it as a graph partitioning problem without setting any assumption on the clusters form. It transforms the given set χ_{lp} into a weighted graph $G = (V, E)$ using some notion of symmetric similarity matrix $A \in \mathbb{R}^{N \times N}$, where each vertex v_i represents $x_{lp\,i}$, and each edge between two vertices v_j and v_i represents the similarity $a_{j,i} \geq 0$ [86]. As mentioned above, it is recommended to use the normalized spectral clustering. Hence, we implemented the NJW[5] algorithm [87], which is detailed in algorithm 3.

Algorithm 3 The NJW spectral clustering algorithm.

Input: The set of data vectors $\chi_{lp} = \{x^{tr}_{lp\,1}, x^{tr}_{lp\,2}, \ldots, x^{tr}_{lp\,N}\}$, the number J of clusters to detect.

Pre-processing:

1: Calculate the similarity matrix A.

2: Calculate the degree matrix D_g, which is a diagonal matrix defined by : $d_{g\;i,i} = \sum_{j=1}^{N} a_{i,j}$.

Spectral representation:

3: Compute the Normalized graph Laplacian matrix $L_{sym} = D_g^{-1/2}(D_g - A)D_g^{-1/2}$[6].

4: Proceed the eigenvalues decomposition of L_{sym} and find the J eigenvectors corresponding to the smallest eigenvalues, arranged in increasing order.

5: Form the matrix U, by stacking the J eigenvectors in columns: $U = [u_1, \ldots, u_J] \in \mathbb{R}^{N \times J}$.

6: Normalize the U's rows to norm 1 in order to get the matrix $U_n \in \mathbb{R}^{N \times J}$, that is, $U_{n_{i,j}} = u_{i,j}/(\sum_j u^2_{i,j})^{1/2}$.

Clustering:

7: Treat each row of U_n, $(u_{n_i})_{i=1,\ldots,N}$, as a data point in $\mathbb{R}^J$, then partition them into J subgroups, $Q_1, \ldots, Q_J$, using the k-means algorithm 4.

8: Attribute the original points $x_{lp\,i}$ to cluster j if and only if row i of the matrix U_n was attributed to cluster j.

Output: Clusters $CL_1, \ldots, CL_J$ with $CL_j = \{i \mid u_{n_i} \in Q_j\}$.

Commonly, identifying the number of clusters J in an optimal manner is the main concern of all clustering algorithms. Generally, with spectral clustering, we find the number J by analyzing the Laplacian matrix eigenvalues that are computed using A

[5] The algorithm name, NJW, is attributed according to the authors' names, that is, Ng, Jordan and Weiss.

[6] The unnormalized graph Laplacian matrix is defined by $L = (D_g - A)$, which refers to the unnormalized spectral clustering.

[7] As an example, we can fix a threshold for the sum of the distances that are computed between the nodes and their respective prototype vectors.

Algorithm 4 The k-means algorithm.

Input: Choosing randomly J different prototype vectors (centroids) $y_1, ..., y_J$ among the data vectors $u_{n_1}, ..., u_{n_N}$.

repeat:

1: Assign each data points u_{n_i} to the closest centroid y_j (in an Euclidean sense). Q_j presents thus the cluster, which contains the objects u_{n_i} that are closest to y_j.

2: Update the new prototype vectors as follows: $y_j = (1/ \mid Q_j \mid) \sum_{u_{n_i} \in Q_j} u_{n_i}$, $\forall j \in [1, J]$.

until an allocated time ends or convergence[7].

and according to the chosen clustering method. In the ideal case, the multiplicity of the eigenvalue 0 equals the number of clusters J. However, this criterion is only valid when the groups are well separated in the graph. In this work, we choose to apply the eigengap heuristic [85], which defines J by finding a drop in the magnitude of Laplacian eigenvalues, $\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$, sorted in increasing order. That is:

$$J = arg \max_{1 \leqslant i \leqslant N} (\lambda_{i+1} - \lambda_i). \tag{4.7}$$

The idea here is to pick the number J in such a way that all the Laplacian eigenvalues $\lambda_1, \ldots, \lambda_J$ are very small compared to λ_{J+1}, which marks relatively a large value.

Regarding the similarity matrix A, we opted for the Gaussian kernel to measure the similarity between the data points $\{x_{lpi}\}$ [87], where σ is a scaling parameter that controls the neighborhoods width:

$$a_{i,j} = \exp(-\frac{\|x_{lpi} - x_{lpj}\|^2}{2\sigma^2}). \tag{4.8}$$

According to [87, Theorem. 2], an appropriate σ can be fixed automatically after repeatedly running the algorithm using a number of values and choosing the one that forms the least distorted partition in the spectral representation space. To determine the appropriate parameter σ, in [85, Section. 8], authors had provided several rules of thumb that are frequently used. For example, the method that we have used states that σ can be chosen to be in the order of nearly the mean distance of a point to its

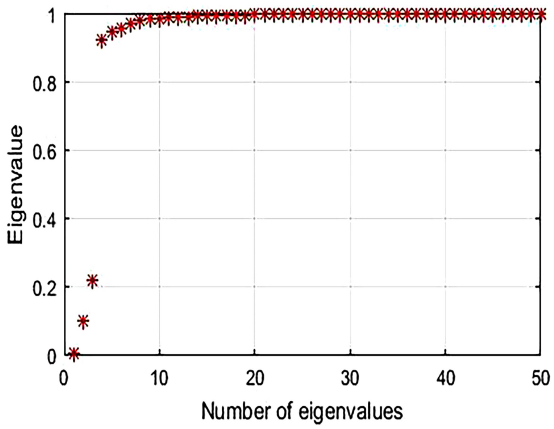


Fig. 4.5. The Laplacian matrix L_{sym} eigenvalues of the generated signal of section 4.3 that are computed using the similarity matrix of (4.8).

k_m^{th} nearest neighbor, where $k_m \sim log(N) + 1$.
Using the first four steps of the aforementioned clustering algorithm 3, Figure 4.5 plots the sorted eigenvalues of the Normalized Laplacian matrix that is computed from the generated signal of the example of section 4.3. Since we used the Gaussian kernel as a similarity matrix, the resulting graph is fully connected, which consists of one connected component. Hence, eigenvalue 0 has multiplicity 1. Clearly, there is a relatively large gap between the 3^{th} and 4^{th} eigenvalue of this trace. According to metric (4.7), the data set contains three clusters, which is well approved.

4.4.2 Sensing and Transmission Schedule

In this part, we determine how we take into account the detected clusters in the representative sensor node selection as well as in the sensing and transmission schedule. Relying on the method of chapter 3, the active node selection process is achieved by considering the inter-spatial correlation between nodes, which can be estimated through (3.2). Different from the previous work of chapter 3 and in order to cover all the clusters, the set $\mathcal{N}_{rep}$ consists of the combination of J subsets, $(\mathcal{N}_{rep_j})_{j=1,\dots,J}$, where $\mathcal{N}_{rep_j}$ includes N_{rep_j} representative nodes picked from cluster CL_j using the

same shared percentage pct_{Nrep}. That is:

$$N_{rep} = \sum_{j=1}^{J} N_{rep_j}, \tag{4.9}$$

where

$$N_{rep_j} = pct_{Nrep}\% \times cl_j. \tag{4.10}$$

In (4.10), if $pct_{Nrep}\% \times cl_j$ is not an integer, we round N_{rep_j} to the nearest integer greater than or equal to the value of that element. Here, the selection of the sets $\mathcal{N}_{rep_j}$ of the clusters' representative nodes is independent from one cluster to another. Hence, the set S_1 appearing in expression (3.2) of chapter 3 is replaced by the set S_1^j, which represents the set of sensor nodes of the cluster CL_j that are not yet selected. Thus, we have:

$$g^* = arg\max_{g \in S_1^j}(m'_g), \tag{4.11}$$

where

$$m'_g = \left(\sum_{i \in S_1^j} \frac{\sigma_{ig}^2}{\sigma_g^2} \right). \tag{4.12}$$

The selection process is the same for the J sets $\mathcal{N}_{rep_j}$. Thus, for each cluster CL_j, according to (4.11), at each iteration $n \in \{1, \ldots, N_{rep_j}\}$, a sensor node $g^*(n)$ is selected and moved from set S_1^j to set S_2^j. Note that S_2^j represents the set of nodes of cluster CL_j that are already chosen during the previous iterations. To proceed with the representative nodes selection procedure, we make use of the learning data matrix $X_{lp} = [x_{lp\,1}^{tr}, x_{lp\,2}^{tr}, \ldots, x_{lp\,N}^{tr}]^{tr} \in \mathbb{R}^{N \times T_{lp}}$ that we partition into J sub-matrices $X_{lp}^j \in \mathbb{R}^{cl_j \times T_{lp}}$, where X_{lp}^j holds data sent by nodes belonging to CL_j. Without loss of generality, for each cluster CL_j and using its corresponding data matrix X_{lp}^j, we perform the steps of the nodes selection process that have been outlined in algorithm 2 of chapter 3, in order to get the set $\mathcal{N}_{rep_j}$, while replacing (3.2) and (3.3) by (4.11) and (4.12) respectively. Here, the proposed data gathering scheme is referred to as the Optimized Cluster-based MC data gathering approach (OCBMC). We denote the OCBMC as the updated version of the Cluster-based MC data gathering approach (CBMC) that has been presented in our paper [25]. Precisely, with the CBMC, the set $\mathcal{N}_{rep}$ of the representative nodes is randomly chosen and with clusters consideration,

whereas, with the OCBMC, the set $\mathcal{N}_{rep}$ of the representative nodes is neatly chosen according to the correlation-based metric (4.11) and with clusters consideration.
Given the example of Figure 4.1, we can note the existence of three detected clusters within the network. We suppose that $pct_{Nrep} = 30$. Thus, 30% of nodes will be selected from each cluster to be active. That is to say that we should pick $N_{rep_1} = 2$ sensors from CL_1, $N_{rep_2} = 1$ sensor from CL_2 and $N_{rep_3} = 3$ sensors from CL_3. That is, in total $N_{rep} = 6$ representative sensors. Based on the correlation among the sensor nodes and using algorithm 2, the obtained subsets are as follows: $\mathcal{N}_{rep_1} = \{13, 1\}$, $\mathcal{N}_{rep_2} = \{9\}$ and $\mathcal{N}_{rep_3} = \{12, 6, 16\}$.

Once the set $\mathcal{N}_{rep}$ of representative sensor nodes is defined, the sink focuses on the sensing and transmitting schedule, Ω_M, by assigning m transmitting source nodes for each time slot t. Obviously, these sensor nodes are picked from the set $\mathcal{N}_{rep}$. Hence, the binary matrix Ω_M consists of N_{rep} $(0, 1)$ binary row vectors and $(N - N_{rep})$ completely zero row vectors. As it has been stated in the previous subsection, in order to ensure the diversity in the delivered data, the m transmitting source nodes are chosen in such a way that we randomly pick, with the same shared percentage pct_m, m_j nodes from each subset $\mathcal{N}_{rep_j}$ corresponding to cluster CL_j. Likewise (4.9) and (4.10), we have:

$$m = \sum_{j=1}^{J} m_j, \tag{4.13}$$

where

$$m_j = pct_m\% \times N_{rep_j}. \tag{4.14}$$

Let us focus again on the example of Figure 4.1. We suppose that $pct_m = 20$. Thus, for each t, 20% of sensors from each subset $\mathcal{N}_{rep_j}$ are randomly designated to deliver their data to the sink. Since the used number N of this example is very small, we end with $m_j = 1$ transmitting source node from each cluster for each t. Note that without enforcing the involvement of all the clusters in the data sensing and transmission process, cluster 2 that contains only sensor 9 could be totally ignored.

To conclude, rather than selecting in a purely random way the measurement locations, as usually used in the conventional MC method, in this part, we presented how to intelligently assign transmitting source nodes that can well represent the network relying on their correlations with the OCBMC.

4.5 The Three-stage MC-based reconstruction approach

After revealing in detail how to select the N_{rep} representative sensor nodes and how to schedule their participation in the data sensing and transmission, we focus, in this section, on how to approximate the entire $N \times T$ data matrix X based on the limited amount of reported readings. Isolating $(N - N_{rep})$ inactive sensor nodes from the sampling and transmission schedule entails the existence of $(N - N_{rep})$ fully empty rows in the received data matrix $M \in \mathbb{R}^{N \times T}$, which impedes the MC technique that is completely unable to estimate the original matrix. Therefore, the use of other complementary interpolation techniques becomes needful. In this context, we develop a structured MC-based recovery framework that is able to ensure the reconstruction of the entire $N \times T$ data matrix X.

4.5.1 Stage 1

Obviously, it is not feasible to directly apply the MC technique with the existence of $(N - N_{rep})$ fully empty rows. Therefore, we have to remove these rows from M. We denote the resultant matrix as $M_{MC} \in \mathbb{R}^{N_{rep} \times T}$, containing the partially delivered readings of the representative sensor nodes. We carry on with the same removal from Ω_M to obtain $\Omega_{MC} \in \mathbb{R}^{N_{rep} \times T}$. Then, making use of the solution introduced in (2.12) or any other method proposed for the MC resolution, we fill the missing entries of M_{MC} that correspond to the non-transmitted data readings of the N_{rep} sensor nodes. As it has been introduced in [42], the threshold parameter τ_{au} roughly equals 100 times the largest singular value of M_{MC}. We denote $X' \in \mathbb{R}^{N_{rep} \times T}$ as the combination of the MC-based estimation and the directly observed data. Finally, we update $X' \in \mathbb{R}^{N \times T}$ by adding the $(N - N_{rep})$ empty rows and placing them in their proper corresponding locations of M.

4.5.2 Stage 2

After filling the random missing readings, remain the $(N - N_{rep})$ completely missing rows that correspond to the inactive sensor nodes. In this phase, we carried on with the spatial pre-interpolation technique of [24, Section. VI], which rebuilds the data of

an empty row using the available data of the neighboring sensor nodes. This method relies on a spatial constraint matrix $H_{sc} \in \mathbb{R}^{N \times N}$, whose computation steps are presented as follows:
1-We start with an identity matrix H_{sc}.
2-For each row $i \in \{1, \ldots, N\}$ of M that corresponds to an inactive node, replace $H_{sc_{(i)}}$ by $Y_{c_{(i)}}$, where $H_{sc_{(i)}}$ and $Y_{c_{(i)}}$ represent respectively the i^{th} row of H_{sc} and the i^{th} row of the topology matrix Y_c, stated in B.1. Then, replace $H_{sc}^{(i)}$ by an all-zero vector, where $H_{sc}^{(i)}$ represents the i^{th} column of H_{sc}. To apply this method, we adjust the 1-hop topology matrix Y_c to the presence of the disjoint clusters in the monitored field, according to B.1, in order to avoid untrustworthy data reconstruction.
3-Finally, the rows of the resulting matrix H_{sc} are normalized in such way that the sum of the elements of a row is 1.
Once H_{sc} is calculated, the spatial pre-interpolation technique can be performed by multiplying H_{sc} by X'. Here, the missing data of an inactive node is obtained using the average of the data readings of its one-hop neighbors.

As mentioned before, the number N_{rep} of the active sensor nodes is very small compared to the total number N, which means that the $(N - N_{rep})$ inactive sensor nodes constitute the preponderant portion of the network. Consequently, there are several IS nodes in the network (having all their neighbors absent). Hence, with the use of the stated topology matrix Y_c, this interpolation technique can achieve the data reconstruction only for the absent sensor nodes, whose neighbors are belonging to the set $\mathcal{N}_{rep}$. We suppose that the network distribution contains N_{Is} *isolated* sensor nodes. Then, the resulting data matrix $X'' \in \mathbb{R}^{N \times T}$, obtained at the end of this stage, i.e. $X'' = H_{sc} \times X'$, still holds N_{Is} empty rows to be recovered (N_{Is} all-zeros rows).

4.5.3 Stage 3

Since the above interpolation technique is limited to recover only a part of the total empty rows (absent nodes), we resort to a second spatial interpolation to rebuild the remaining part of the empty rows (*isolated* nodes). Benefiting once again from the spatial dependency among the sensor nodes, we fill the remaining empty rows using

the following minimization problem:

$$\underset{\widehat{X}\in\mathbb{R}^{N\times T}}{\text{minimize}}\,(fac_1\times\|\widehat{X}-X''\|_F^2+fac_2\times\|S\times\widehat{X}\|_F^2), \tag{4.15}$$

where S represents a spatial constraint matrix, whose computation steps will be detailed hereafter, fac_1 and fac_2 are two tuning parameters and $\widehat{X}\in\mathbb{R}^{N\times T}$ is the final reconstructed data matrix. The resolution of this optimization problem can be easily accomplished using the semidefinite programming (SDP). To solve (4.15) and obtain $\widehat{X}$, we opted for the CVX package [88], implemented in Matlab, as an advanced convex programming solver.

In this equation, the matrix $S\in\mathbb{R}^{N\times N}$ relatively reflects our knowledge about the spatial structure inherent in the data since it is computed based on the learning data matrix $X_{lp}\in\mathbb{R}^{N\times T_{lp}}$. This spatial matrix expresses the similarities between the sensor nodes readings. Suitably, we use the Euclidean distance as a distance function, computed in the data domain of the sensor nodes, to model the similarity between the rows of X_{lp}. Indeed, the smaller the distance between two rows, the closer they are. Below are the steps to determine S [44]:

1-We initiate these steps with an all-zeros matrix S.

2-The similarity between the rows in X_{lp} is not evident as the ordering of the sensor nodes indexes in X_{lp} is arbitrary. Thus, for each row i of X_{lp}, we search for the set j_i' of indexes of the K closest rows to i, that is, $j_i'=\{j_k\neq i\mid k=1,...,K\}$.

3-Assuming that the row i can be approximated through the linear combination of the rows of set j_i', we perform the linear regression to compute the weight vector $W_i=[w_i(1),\ldots,w_i(K)]\in\mathbb{R}^{1\times K}$ through the following equation:

$$W_i=X_{lp}(i,:)\times X_{lp}(j_i',:)^T\times[X_{lp}(j_i',:)\times X_{lp}(j_i',:)^T]^{-1}. \tag{4.16}$$

4-Finally, we assign 1 to $S(i,i)$ and $-w_i(k)$ to $S(i,j_k)$.

As soon as these steps have been carried out for all the rows i, we obtain the matrix S, with which we interpolate $\widehat{X}$ as in (4.15).

Now, remains the last adjustment to realize, that is, the scaling of the two parameters, fac_1 and fac_2 of (4.15). The regularization parameters fac_1 and fac_2 are introduced

in order to establish a trade-off between a close fit to the matrix X'' and the intention of fulfilling the N_{Is} remaining empty rows using S. It will be shown through simulations that adjusting these parameters nicely improves the reconstruction performance [44], and the founded values of fac_1 and fac_2 are independent of the size of the matrix (N and T).

Let us focus again on the example of Figure 4.1. The dotted lines refer to the neighborhood relation between sensors. As we can see, the sensors $\{5, 8, 10, 11, 14\}$ are each linked at least to a representative sensor. Thus, their data readings can be easily recovered through the spatial pre-interpolation method of stage 2. However, the nodes $\{2, 3, 4, 7, 15\}$ are considered as *isolated* from the network. Thus, their readings are recovered thanks to the minimization (4.15) of stage 3.

4.6 Numerical Results

In this section, we compare the performance of our proposed structured approach to that of a benchmark scheme, which was designed basically on what was proposed in [24] and in line with our scenario requirements. Indeed, at the end of their work, Xie et al. considered in [24] that there is a small number of empty rows in M, that is, for $N = 196$, 14 data rows were missing, namely 7% of N (i.e. 93% of N of representative sensors). As we have already stated at the beginning of this chapter, treating an important number of missing data rows has not been the main focus of their work. Thus, their proposed approach has not taken into account the existence of the *isolated* nodes in the network. In fact, they focused basically on the existence of successive missing or corrupted entries in the received data matrix M. However, to the best of our knowledge, this is the unique approach that has treated a similar case using MC, and with which we can compare our approach in the first part of this section. Then, in the second part, we try to evaluate separately the benefits of each building block of the proposed approach, namely:

- Involving all the detected clusters equitably in the sampling process using (4.9, 4.10) and (4.13, 4.14).
- Selecting the representative sensor nodes using algorithm 2 with (4.11) and (4.12).

- Adding the minimization (4.15) to the reconstruction pattern.

Making use of the multi-Gaussian signal model of section 4.3, we perform our structured approach over different scenarios to illustrate the impact of these aforementioned techniques on the interpolation accuracy of the data matrix. To measure the reconstruction error, we opted for the following metrics, where X and $\widehat{X}$ represent respectively the initial raw data matrix and the reconstructed one:

1-$NMAE_{tot}$: The Normalized Mean Absolute Error on all missing entries:

$$NMAE_{tot} = \frac{\sum_{i,t:\Omega_M(i,t)=0} |X(i,t) - \widehat{X}(i,t)|}{\sum_{i,t:\Omega_M(i,t)=0} |X(i,t)|}. \tag{4.17}$$

2-$NMAE_{MC}$: The Normalized Mean Absolute Error on the partially missing entries, which correspond to the non-transmitted readings of the N_{rep} representative nodes:

$$NMAE_{MC} = \frac{\sum_{i,t:(i,t)\in\Omega_{mc}} |X(i,t) - \widehat{X}(i,t)|}{\sum_{i,t:(i,t)\in\Omega_{mc}} |X(i,t)|}, \tag{4.18}$$

where Ω_{mc} is the set of indexes of the partially missing entries, found in the received data matrix $M \in \mathbb{R}^{N\times T}$. This metric measures the error ratio following the 1^{rst} stage of the reconstruction pattern.

3-$NMAE_{ER}$: The Normalized Mean Absolute Error on the missing entries of the fully empty rows, which correspond to the inactive sensor nodes readings:

$$NMAE_{ER} = \frac{\sum_{i,t:i\in\Omega_{ER}} |X(i,t) - \widehat{X}(i,t)|}{\sum_{i,t:i\in\Omega_{ER}} |X(i,t)|}, \tag{4.19}$$

where Ω_{ER} is the set of indexes of the $(N - N_{rep})$ empty rows, found in the received data matrix $M \in \mathbb{R}^{N\times T}$. This metric measures the error ratio following the 2^{nd} and the 3^{rd} stages of the reconstruction pattern.

4-CR: The Compression Ratio:

$$CR = \frac{N \times T - card(\Omega)}{N \times T}, \tag{4.20}$$

where $\Omega = \{(i,t) \mid \Omega_M(i,t) = 1\}$. Hence, $card(\Omega)$ denotes the number of observed

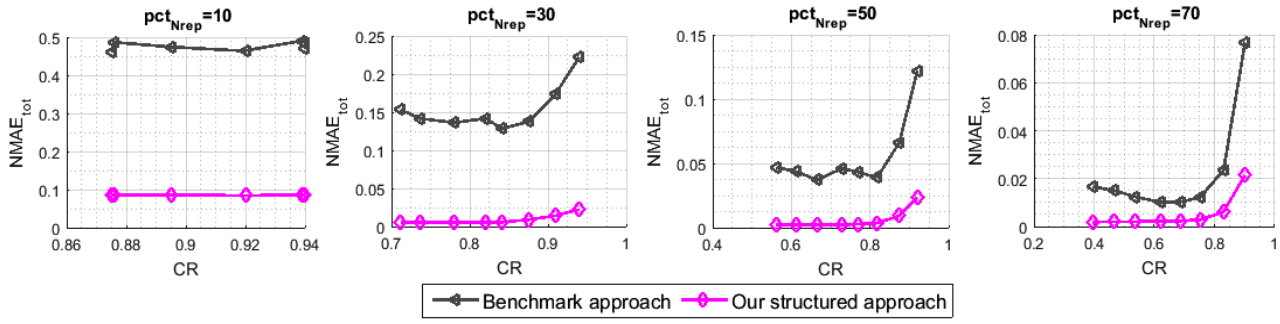


Fig. 4.6. $NMAE_{tot}$ **for the proposed technique and for the Benchmark.**

entries in M.

To assess the proposed approach under different CRs, we vary pct_{Nrep} from 10 to 80, and for each given pct_{Nrep}, we vary pct_m from 10 to 80. It is obvious that the range of the values of CR depends on the value assigned to pct_{Nrep}. The larger pct_{Nrep}, the higher CR range can be used. Note that we are mainly interested in the small values of pct_{Nrep} and pct_m since we are considering the high loss scenarios.
Specifically, we consider that $N = 50$ sensor nodes are randomly distributed in a square observation area of size $100m \times 100m$, and we monitor the WSN during $T = 100$ time slots. To perform the minimization (4.15) of stage 3, the parameters that we have used during all the simulations have been determined empirically, and are given as follows; $K = 5$, $fac_1 = 10^{-13}$ and $fac_2 = 1$. To find out how we have chosen these tuning parameters, see B.3 in appendix B.

To begin, we implement a benchmark approach based on what was proposed in [24]. The sampling pattern of this approach consists in choosing the set $\mathcal{N}_{rep}$ of representative sensor nodes in a purely random way, which is exactly the same as randomly selecting the empty rows. Likewise, for each time slot t, m nodes are uniformly selected from the set $\mathcal{N}_{rep}$ to deliver their readings to the sink. Here, neither the selection of the representative sensors nor the selection of the transmitting source ones takes into account the detected clusters. As for the reconstruction pattern, to obtain the final recovered data matrix $\widehat{X}$, this approach performs the MC then the spatial pre-interpolation. The temporal pre-interpolation was omitted since we don't consider the existence of empty columns in the observed data matrix M[8]. In Figure 4.6, we

[8]This is not the case with our scenario since, at every t, we ensure the transmission of m readings sensed in different m locations.

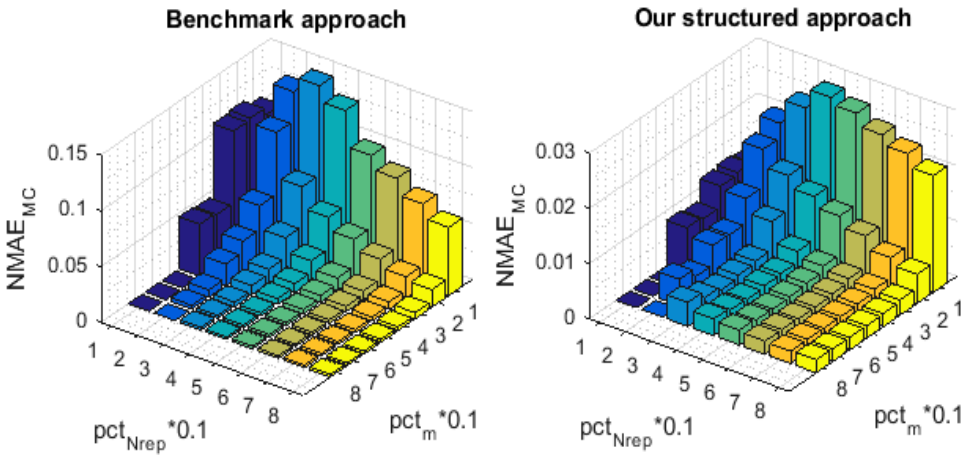


Fig. 4.7. $NMAE_{MC}$ **for the proposed technique and for the Benchmark.**

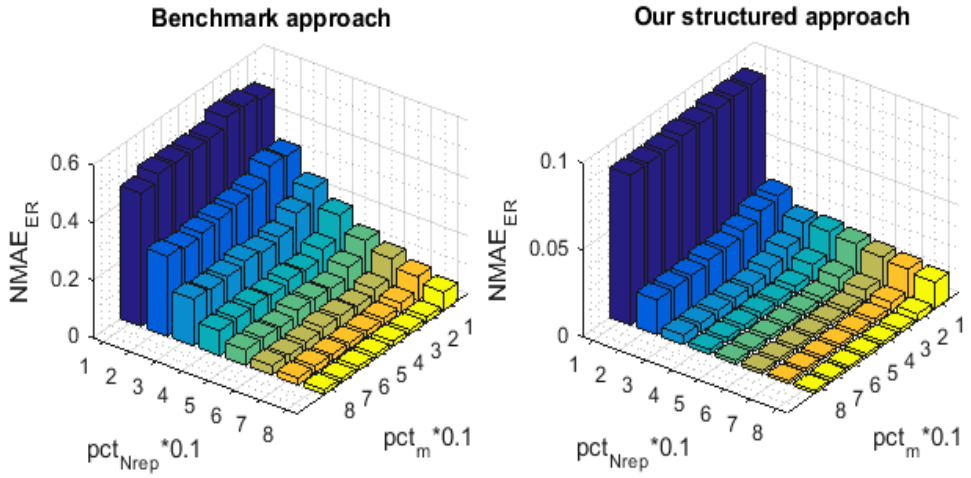


Fig. 4.8. $NMAE_{ER}$ **for the proposed technique and for the Benchmark.**

have measured the $NMAE_{tot}$ with respect to the variation of CR, namely pct_m, for different values of pct_{Nrep}. As we can note from the plots, our approach distinctly outperforms the benchmark one across the entire ranges of CR. We are able to go up to 90% of missing rows ($pct_{Nrep} = 10$) with an interesting reconstruction performance, $NMAE_{tot}$ of about 0.08, while the benchmark technique yields an $NMAE_{tot}$ of $[0.46, 0.5]$. Figures 4.7 and 4.8 illustrate the 3-D bar graph of respectively the $NMAE_{MC}$ and the $NMAE_{ER}$ values with the variation of pct_{Nrep} and pct_m. For the convenience of comparison, we have implemented the $NMAE_{MC}$ and the $NMAE_{ER}$

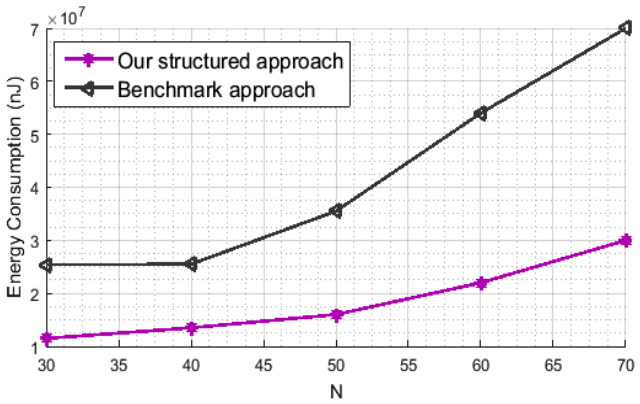


Fig. 4.9. Energy consumption for the proposed technique and for the Benchmark.

in order to separate the error ratios and demonstrate the recovery performance enhancement achieved by our proposed approach on respectively the partially and the fully missing readings.

Note that the considered framework extremely reduces the overall network energy consumption since we only use a small set of representative sensors for the data transmission. Furthermore, compared to the benchmark approach, the proposed one can further improve the sensors lifetime. In fact, for a given $NMAE_{tot}$ target of 0.02 and $pct_{Nrep} = 60$, we compute the energy consumption during the T time slots for the both compared approaches depending on the number N of sensors. In this simulation, as in chapter 3, we consider that two nodes i and j can directly communicate with each other, without the need for relaying, only if the Euclidean distance $dst_{i,j}$ between them is within some transmission radius (r) that scales with $\ominus(\sqrt{logN/N})$, and to route the data towards the sink node, we perform the shortest path tree computed by Dijkstra algorithm. In order to compute the energy consumption during data transmission, the following model is used [89]:

$$\begin{cases} E_{Tx}(L, dst_{i,j}) = E_{elec} \times L + \varepsilon_{amp} \times L \times dst_{i,j}^2 \\ E_{Rx}(L) = E_{elec} \times L, \end{cases} \tag{4.21}$$

where $E_{Tx}(L, dst_{i,j})$ and $E_{Rx}(L)$ represent respectively the amount of energy consumed by a specific node i, to deliver and receive an L-bit packet through a distance of length $dst_{i,j}$. In (4.21), E_{elec} is the energy required by the transceiver circuitry at the

sender or the receiver and ε_{amp} is the energy consumed by the transmitter amplifier. Hence, depending on the distance dst between the transmitter and the receiver, the total energy cost for forwarding L bits of data is $E_{Tx}(L, dst)+E_{Rx}(L)$. Regarding the parameters setting, $L = 120$ bits [21], $E_{elec} = 50\ nJ/bit$ and $\varepsilon_{amp} = 100\ pJ/bit/m^2$ [89]. Figure 4.9 illustrates the energy consumption for the proposed framework as well as for the benchmark one. Indeed, our approach requires far less sensor nodes' readings, consequently much less energy consumption, to achieve the same reconstruction performance.

Let us focus now on the benefits of the clusters selection. We show that taking into account the detected clusters during the representative nodes selection process as well as during the assignment of the sensing and transmitting schedule significantly ameliorates the data recovery performance. Thus, we compare our approach to another one, for which we proceed regardless the existence of the different clusters. The set $\mathcal{N}_{rep}$ of representative sensor nodes is selected according to (3.2) and (3.3) instead of (4.11) and (4.12), i.e. the spatial correlation criteria is present during the nodes selection process. Nevertheless, we do not have equitable representation of the different regions that compose the whole network. Withal, for each t, the m transmitting source nodes are picked from the set $\mathcal{N}_{rep}$ in a purely random way to sense then deliver their data readings, i.e. $m = pct_m\% \times N_{rep}$ instead of (4.13) and (4.14). To recover the received data matrix, both algorithms apply the three-stage MC-based reconstruction pattern of section 4.5. Figure 4.10 illustrates the 3-D bar graph of the $NMAE_{tot}$ values with the variation of pct_{Nrep} and pct_m. This simulation shows how curiously interesting the clusters consideration is. The barres depict that our approach provides a considerable improvement in terms of $NMAE_{tot}$ compared to the algorithm of comparison, especially in the high compression ratios, i.e. when the number of transmitting source nodes is very limited. Note that without enforcing the involvement of all the clusters in the data sensing and transmission process, sensor nodes belonging to the small clusters could be totally ignored, which gravely deteriorates the recovery process. In Figures 4.11 and 4.12, we have measured respectively the $NMAE_{MC}$ and the $NMAE_{ER}$ with respect to the variation of CR, namely pct_m, for different values of pct_{Nrep}. Figures 4.11 and 4.12 highlight the effect of the introduced block on the data recovery of respectively the representative nodes and the inactive nodes readings.

Although both techniques apply the same MC resolution method, the $NMAE_{MC}$ of

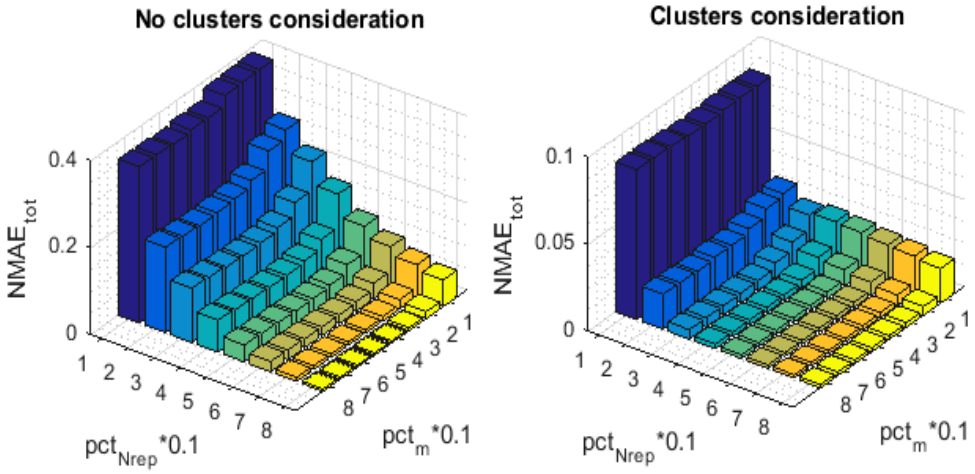


Fig. 4.10. $NMAE_{tot}$ **with and without clusters consideration.**

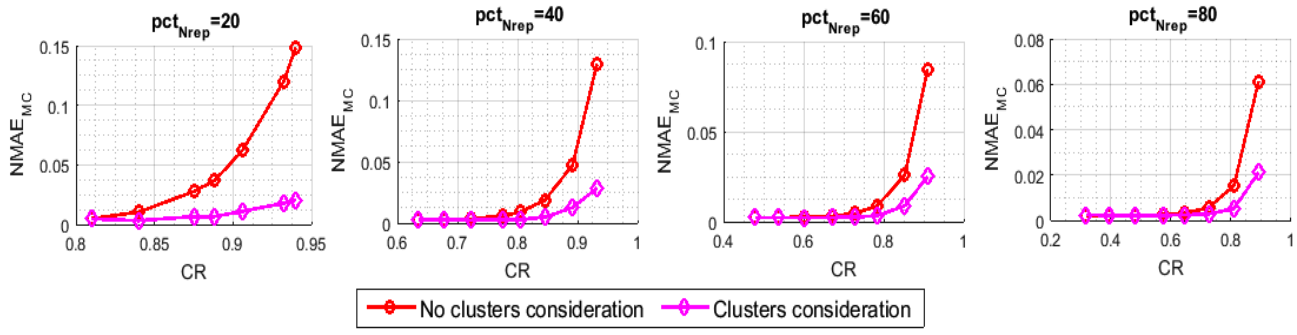


Fig. 4.11. $NMAE_{MC}$ **with and without clusters consideration.**

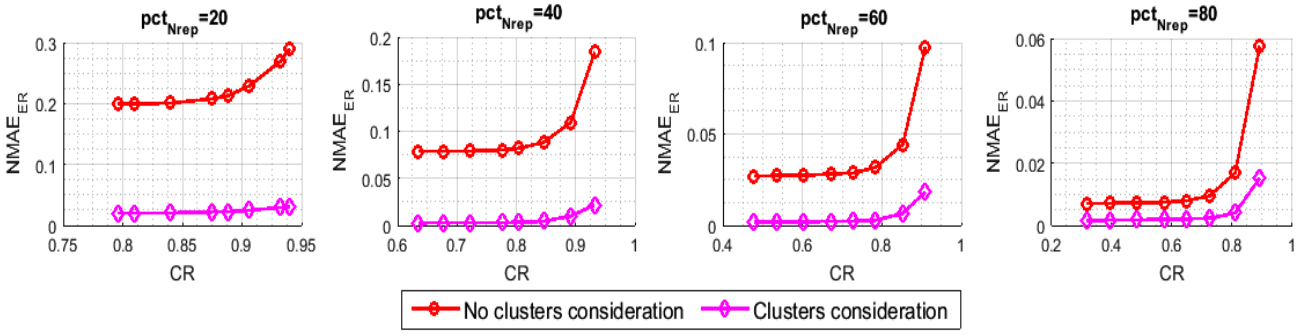


Fig. 4.12. $NMAE_{ER}$ **with and without clusters consideration.**

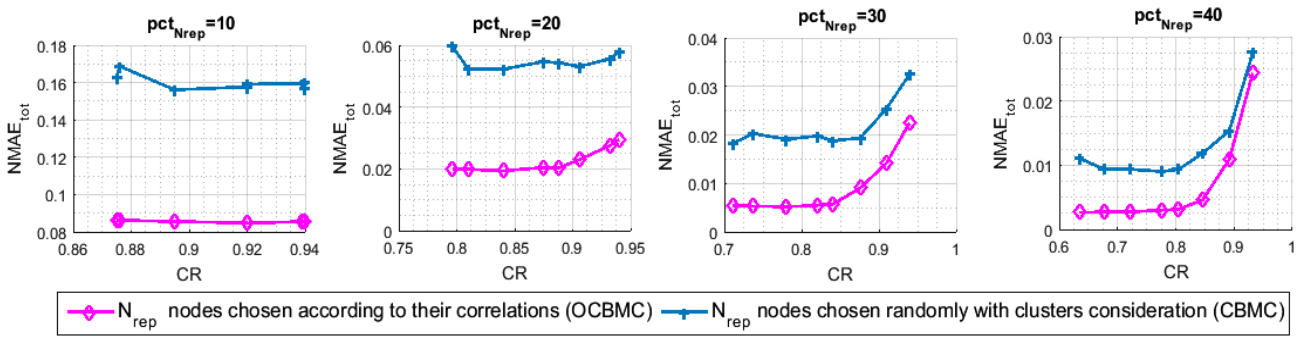


Fig. 4.13. The impact of the representative node selection technique on the $NMAE_{tot}$**.**

our approach is much lower than that of the benchmark, especially for the small values of pct_m. The $NMAE_{ER}$ also seems to be heavily affected, despite the fact that the clusters consideration, at the base, targets only the first stage of the reconstruction pattern, which is the MC resolution. For example, with $(pct_{Nrep} = 20, pct_m = 10)$, $(pct_{Nrep} = 40, pct_m = 10)$ and $(pct_{Nrep} = 60, pct_m = 10)$ we can reach an improvement, on the $NMAE_{ER}$, respectively of 89.619%, 88.587% and 81.443%, when we enforce the involvement of all the clusters in the data sensing and transmission.

The next scenario aims to prove the importance of neatly selecting the N_{rep} representative nodes. Making use of the spatial correlation in the selection process, these nodes are selected under the criterion of having the best representation of the whole network. To investigate the efficiency of the proposed selection process, we compare our algorithm to another one that selects its representative nodes randomly. However, in order to be comparable, this one takes into account the existing clusters when selecting its representative nodes. Hence, the set $\mathcal{N}_{rep}$ of representative nodes consists of the combination of J subsets, $(\mathcal{N}_{rep_j})_{j=1,\dots,J}$, where $\mathcal{N}_{rep_j}$ includes N_{rep_j} representative nodes selected randomly from cluster CL_j using the same shared percentage pct_{Nrep}, where $N_{rep} = \sum_{j=1}^{J} N_{rep_j}$ and $N_{rep_j} = pct_{Nrep}\% \times cl_j$. As described in 4.4.2 and according to (4.13) and (4.14), both algorithms design their sensing and transmitting schedules, $\Omega_M \in \mathbb{R}^{N \times T}$, based on their selected sets $\mathcal{N}_{rep}$ of representative nodes. To recover the received data matrix, both algorithms apply the three-stage MC-based reconstruction pattern of section 4.5. Typically, the algorithm of comparison represents that proposed in our paper [25] (i.e the CBMC). As we have stated in section 4.4.2, the use of a selection cost function (4.11) represents a developed update

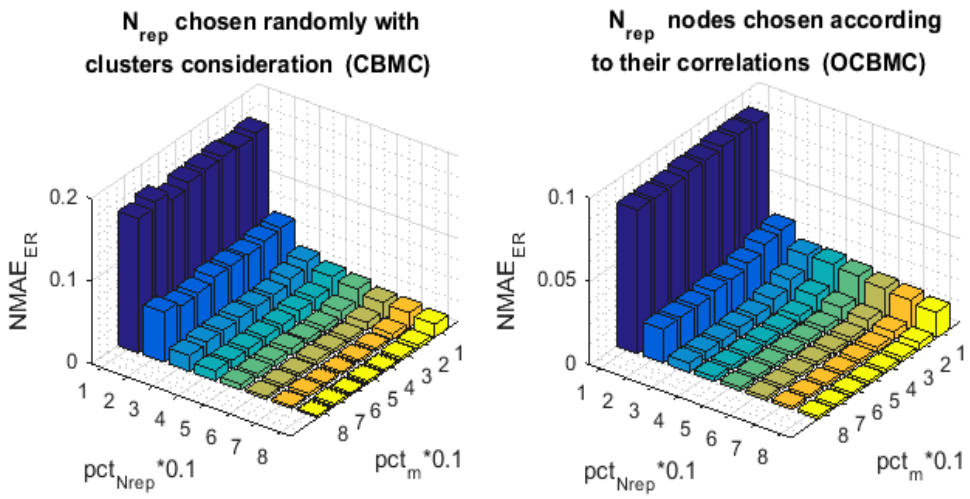


Fig. 4.14. The impact of the representative node selection technique on the $NMAE_{ER}$.

or an improvement to the proposed approach of [25]. The results of this simulation are depicted in Figures 4.13 and 4.14. Figure 4.13 illustrates the $NMAE_{tot}$, and as we can see, compared to the random selection process, the selection scheme of algorithm 2 with (4.11) and (4.12) (i.e the OCBMC) provides considerable improvement in term of $NMAE_{tot}$ across the entire ranges of CRs. The gap between the two curves decreases as we increase the number N_{rep} of representative nodes, namely pct_{Nrep}, since we decrease the probability of choosing different sets $\mathcal{N}_{rep}$. Nevertheless, as we have already stated, these cases are not of prime interest for us. Let us focus now on Figure 4.14 that highlights the $NMAE_{ER}$ to reveal the impact of our selection process on the reconstruction performance of the empty rows. Expectedly, we find that the $NMAE_{ER}$ is sensitive to the used selection method, which confirms the aforementioned hypothesis. That is, in order to guarantee an accurate reconstruction for the inactive nodes missing data, a great care must be taken when selecting the set $\mathcal{N}_{rep}$ of representative nodes.

The third simulation highlights the benefit of the 3^{rd} stage of the proposed reconstruction pattern. We compare our algorithm to the one that uses only the first two stages of section 4.5 to get its final recovered data matrix $\widehat{X}$. Following the same logic of the previous experiences, in order to be comparable, we use the sampling pattern of section 4.4 with both simulated algorithms, which yields the same set $\mathcal{N}_{rep}$ of rep-

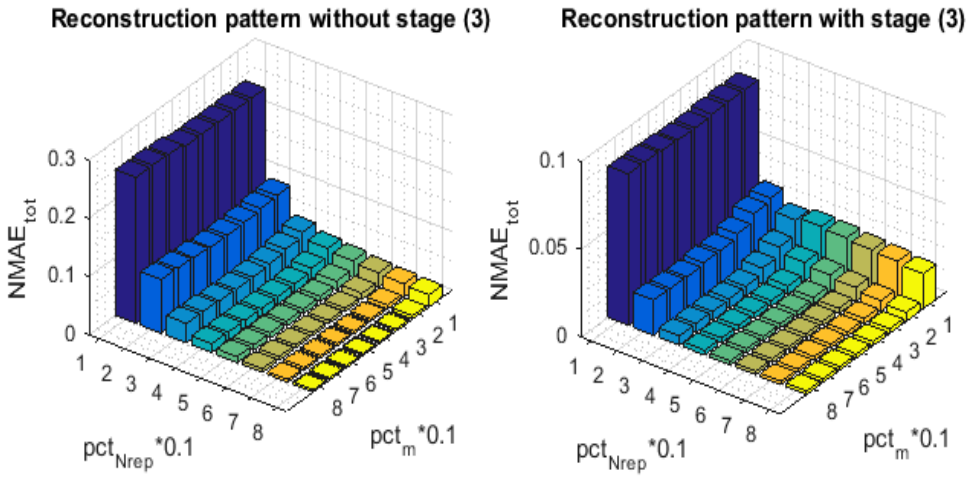


Fig. 4.15. The impact of the spatial interpolation technique on the $NMAE_{tot}$.

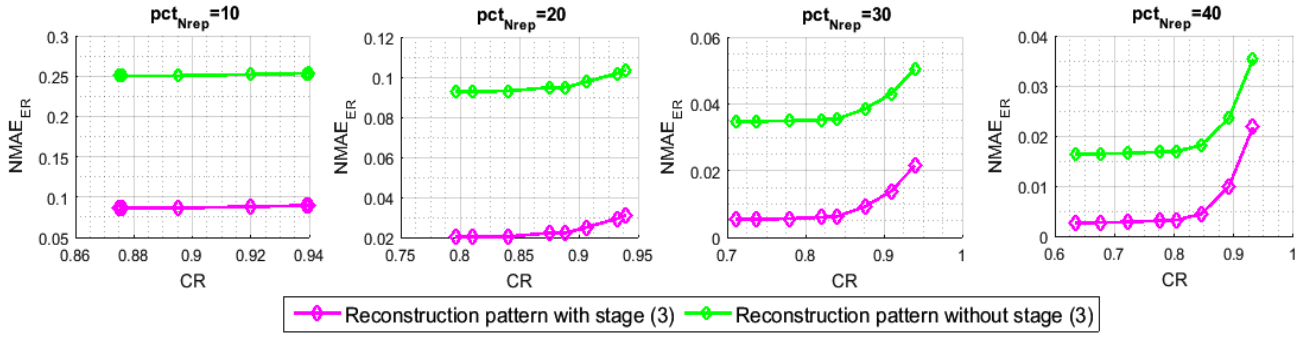


Fig. 4.16. The impact of the spatial interpolation technique on the $NMAE_{ER}$.

resentative nodes and consequently the same set of inactive nodes. Noticeably, we can detect a considerable gap in terms of $NMAE_{tot}$ between the barres of Figure 4.15. This difference for all the pct_{Nrep} values comes from the non-reconstructed readings of the N_{Is} *isolated* nodes with the algorithm of comparison. Since we simulated the same network with the same sensor nodes neighboring, the set of the N_{Is} *isolated* nodes is the same for both of the compared algorithms. Figure 4.16, which depicts the $NMAE_{ER}$ for both approaches, illustrates that we can reduce the reconstruction error of the empty rows up to 65.079% for $(pct_{Nrep} = 10, pct_m = 40)$, 76.842% for $(pct_{Nrep} = 20, pct_m = 40)$, 82.857% for $(pct_{Nrep} = 30, pct_m = 40)$ and 82.353% for $(pct_{Nrep} = 40, pct_m = 40)$, when we apply the minimization (4.15). These results show that the number of *isolated* nodes is important for the small pct_{Nrep} values. Hence,

adding a third interpolation technique, as our proposed minimization (4.15), becomes heavily needed. Otherwise, we can end up with a data matrix, which is almost half built, even less.

4.7 Conclusion

In this chapter, we have proposed to let a significant number of sensor nodes remain idle. Then, relying on a novel MC-based reconstruction framework, we recover their readings based on the received ones. The strength of our approach lies in its integration or inclusivity for both the compression and the reconstruction patterns. For the sampling part, by making use of the inter-spatial correlation feature, we have used a cluster-based strategy that neatly selects a restricted number of representative sensor nodes from each cluster in order to efficiently afterwards schedule where and when to sense the field. As for the reconstruction part, by taking advantage of the readings similarities in the WSNs, we propose an optimization technique that is annexed to the MC resolution. This method, positioned in the third stage of the recovery operation, guarantees the reconstruction of all the empty rows corresponding to the inactive sensor nodes. Altogether, these techniques succeed in handling the aforementioned high loss scenario. We have obtained satisfactory results proving the efficiency and the robustness of the proposed techniques as well as the whole unified approach. The results, obtained with the multi-Gaussian generated signal, outperform those of all the state-of-art techniques. They revealed that we are able to go up to 90% of missing rows (i.e. only 10% of N of representative sensor nodes), while we still achieve interesting data reconstruction accuracy by giving a $NMAE_{tot}$ of about 0.08 compared to the benchmark one, which is still within the range of $[0.46, 0.5]$.

Chapter 5

The Energy-Aware Matrix Completion based Data Gathering Scheme

Contents

5.1 Introduction

In this chapter we carry on with the twofold data compression scenario that has been addressed in the previous chapter 4 with the OCBMC. Reducing the amount of sensing data can indeed minimize the power consumption of the network and save its energy. Nevertheless, it is not sufficient since it does not necessarily alleviate the problem of energy load imbalance between nodes. Indeed, depending on the events to be monitored, even though the representative sensors may change from one detection period to another, the signal in most WSNs is time-stationary. Hence, the set of selected representative nodes can remain the same for many successive sensing periods. To avoid the overcharge that may occur over some sensor nodes, and thus their fast death, the representative nodes should be changed from a detection period to another. In addition, in the multi-hop WSNs, data packets that are generated from the transmitting source nodes should be relayed via intermediate nodes to be routed to the sink. Accordingly, nodes around the sink would exhaust their batteries faster as they carry heavier traffic loads than the border nodes, causing the problem of energy hole. In this case, even if the rest of nodes still hold sufficient energy levels, communication with the sink will be cut off leading to the end the network lifespan. To overcome the issue of uneven energy depletion phenomenon, in addition to the correlation, we have incorporated the sensors' residual energies in the representative node selection function with the proposed Energy-Aware MC-based data gathering approach (EAMC). It is noteworthy that taking into account the residual energy in the node selection process is related to the type of application one wants to perform. To this end, we have evaluated our selection strategy under different scenarios and network topologies while presenting for each one the adequate energy-aware metric. More specifically, our main contributions in this chapter are given as follows:

- As a sequel of chapter 4, in this chapter, we focus on the node selection process taking into account the reconstruction quality as well as the energy efficiency. In addition to the correlation, we have incorporated the sensors' residual energies in the representative node selection function to develop different energy-aware cost selection functions for the EAMC. The proposed combined metrics have been introduced in order to systematically maintain a load balancing among nodes and thus maximize the network lifetime, while still achieving a low data

reconstruction error.

- Different topologies and scenarios have been assessed under the adequate energy-aware proposed metric. Indeed, in the star topology networks, where communication with the sink is direct, choosing a node to be a representative one according to its residual energy in order to improve the network energy utilization is sufficient. However, in the mesh topology networks, where routing schemes must be applied and data is forwarded via relaying nodes, the entire route should be assessed. In addition to the correlation, a node can be chosen to be a representative one if there is no depleted relaying node in its route.
- In this chapter, we target to minimize the energy consumption and extend the network lifespan through nodes energy load balancing, while, at the same time, ensuring a sufficiently good quality of data reconstruction. In the numerical results section, we have studied the trade-off between the data recovery error and the network lifetime for all the investigated scenarios.
- The assumption that the energy consumed in the data acquisition is much lower than that consumed in radio communications does not hold for a number of practical applications, such as the gas sensors which are considered as power greedy sensors [7]. Therefore, in this chapter, we have assessed our approach under both sensor nodes types, the ordinary sensors (low sensing power sensors) and the power greedy ones.

The chapter is organized as follows. The next section is devoted to state the preliminary and the energy consumption system model. In section 5.3, we present the proposed energy-aware data gathering strategy under different scenarios. Then, in order to evaluate the performance of the proposed scheme, we carry out, in section 5.4, with various simulations, where we vary the cost selection function, the addressed scenario and the type of the deployed sensor nodes. Finally, we conclude the work in section 5.5.

5.2 Preliminary and Energy Consumption Model

5.2.1 Preliminary

In this work, we keep using the technique proposed in chapter 4, section 4.5, i.e the three-stage MC-based reconstruction approach. The sink node applies this technique to recover the entire data matrix $X \in \mathbb{R}^{N \times T}$, after receiving a partly empty matrix $M \in \mathbb{R}^{N \times T}$, where N denotes the number of deployed sensor nodes, and T designates the number of time slots t composing the detection period. The three-stage MC-based reconstruction framework is considered as a data recovery building block for all the data gathering schemes that will be introduced in section 5.3. Moreover, note that we keep using notations used along the previous chapter such as those related to the representative nodes, transmitting source nodes and clusters consideration.

5.2.2 The Energy Consumption Model

Generally, a sensor node consumes the energy of its battery in three operations that are communications (i.e. both data transmission and reception), data sensing and data processing.

Since with the MC method, there is no on-sensor computation, and data is directly sub-sampled in the compressed form (i.e. the data $x_{i,t}$ is available only if a location i is chosen to be sensed in the time slot t), we assume here that there is no energy consumed in data processing. Moreover, the high energy-intensive reconstruction algorithm is executed at the sink node, which is free of energy constraint and whose energy consumption does not be included in the network overall energy consumption.

Regarding the transmission and reception activities, we consider the model (4.21) of the previous chapter in which we differentiate the amount of energy consumed by the transceiver circuitry at the sender, i.e $E_{elec-tr}$, to that consumed by the transceiver circuitry at the receiver i.e $E_{elec-rc}$:

$$\begin{cases} E_{Tx}(L, dst_{i,j}) = E_{elec-tr} \times L + \varepsilon_{amp} \times L \times dst_{i,j}^2 \\ E_{Rx}(L) = E_{elec-rc} \times L, \end{cases} \tag{5.1}$$

To monitor the network area and sense the data field, we have used the following expression to compute the energy dissipation by a sensor node when performing the sensing operation for L bit packet [8]:

$$E_{sens}(L) = L \times V_{sup} \times I_{sens} \times T_{sens}, \tag{5.2}$$

where V_{sup} is the supply voltage, I_{sens} is the total current required for the data sensing operation, and T_{sens} denotes the time duration allowed to a sensor node for data sensing.

5.3 Our proposed data gathering scheme

In this section, we present how the energy constraint can be jointly considered with the correlation criteria in the active node selection process in order to maintain a load balancing among nodes and maximize the network lifetime, while still achieving a low data reconstruction error. Since the performance usually vary with the network configurations, we differentiate, in this section, the proposed energy-aware cost functions for the representative node selection according to the given network topologies.

Usually, nodes are randomly scattered in the area to be monitored, without any infrastructure, leading to the existence of different network topologies, which are determined according to the nodes' locations and the connections between them and the sink node. Different topologies may exist, in the WSNs, and vary with the kind of application one wants to proceed. In the sequel, we consider the frequently used topologies, which are the star and the tree/mesh topologies with the twofold addressed scenario.

5.3.1 Single-Hop Star Topology

The star topology networks are single-hop systems [90] since all nodes operate as terminal devices and directly communicate with a centralized communication server. This type of architecture is generally used in wireless micro sensor networks as the covered area is, most of the time, small and limited by the communication range of the end nodes. As we have previously stated, the first step in the network sampling proceeding is to partition nodes into J disjoint clusters. Performing this step is of

prime importance to reach an adaptive and overall representation for the whole monitored area, and thus a more efficient data sampling. Benefiting from the dependency among nodes, the aforementioned representative node selection strategy, using algorithm 2 with (4.11) and (4.12), targets to achieve a better data sampling quality and hence a much lower data reconstruction error at the sink node, despite the limited number of reported data readings with the addressed twofold data compression scenario. However, there is still a crucial factor that cannot be overlooked at all, and must be cautiously taken into consideration, which is the network lifespan and energy load balancing between nodes. Indeed, depending on the events to be monitored, even though the set of representative sensors may change from one detection period T to another, the signal in most WSNs is time-stationary. Hence, the set $\mathcal{N}_{rep}$ of selected representative nodes can remain the same for many successive detection periods. To avoid the overcharge that may occur over some continuously operating sensor nodes and thus the fast depletion of their batteries, the active node selection process should take into account not only correlation between sensor nodes but also their residual energies. Accordingly, we incorporate in (4.11) the fraction of the sensor residual energy, as a complementary factor, in order to choose the sensor nodes that can well represent the network and at the same time hold the highest residual energy. Precisely, for a given sensor node $g \in S_1^j$, the trade-off between its informative value $m_g^{'}$, computed in (4.12), and its residual energy with regard to the other sensors' residual energies, Ef_{resd_g}, is achieved through a multiplication of the two considered factors. Thereby, (4.11) is replaced by (5.3) for our EAMC approach:

$$g^* = arg \max_{g \in S_1^j} \left(m_g' \times Ef_{resd_g} \right), \tag{5.3}$$

where

$$Ef_{resd_g} = \frac{E_g}{\sum_{i \in S_1^j} E_i}. \tag{5.4}$$

Thus, the EAMC represents an update of the OCBMC. The unique difference here is that, with the OCBMC scheme, the set $\mathcal{N}_{rep_j}$ is selected from cluster CL_j passing through the correlation-based cost function (4.11), whereas, with the EAMC, this set is selected from CL_j according to the combined energy-aware and correlation-based metric (5.3).

Performing (5.3) means that we attempt to choose the sensor node carrying the maximum value of the combined metric ($m'_g \times Ef_{resd_g}$). Here, multiplying the two addressed factors aggregates them into a one single entity, and it is analogous to computing the needed correlation per unit of energy. In other words, this operation makes the relation between the two factors fusional. If one of them is weak it will automatically weaken the other, and the carrier sensor node will not be chosen. Since the residual energy of the operating nodes decreases from one detection period to another, the metrics $(m'_g \times Ef_{resd_g})_{g \in S_1^j}$ vary and the representative nodes will be selected efficiently, according to the available energy in their batteries.
In order to determine the set $\mathcal{N}_{rep_j}$ of the EAMC, we perform the same steps of the nodes selection process that have been outlined in algorithm 2, while replacing only the metric (4.11) by the metric (5.3).

5.3.2 Multi-Hop Mesh Topology

Compared to the star topologies, the mesh network does not suffer from the limited scalability. Thus, much wider area can be covered and monitored thanks to the multi-hop transmissions. In this type of networks, several routes may exist between sensor nodes and the sink, and most of the time the network software chooses the shortest one for data delivery. To forward the data towards the sink, we opted for the shortest path tree, implemented with Dijkstra algorithm. Note that the routing protocol to use is not the main focus of this work since our aim is to achieve energy load balancing between nodes and reach a higher lifetime for the network with the already established routes. Updating the paths systematically according to the remaining energy levels in order to further prolong the network lifetime is left as a perspective for future works. A more detailed discussion on this point is afforded in the last chapter 6, section 6.2.

In light of the importance of energy utilization enhancement, as far as the size of these networks gets bigger and the diameter of the covered area gets larger, the problem of uneven energy depletion aggravates and gets worse. In fact, data packets, which are generated by the transmitting source nodes, have to be relayed through intermediate nodes to be finally routed to the sink. Accordingly, nodes that are close to the sink are susceptible to carry much heavier traffic loads than nodes of the outer-regions. Consequently, they would speedily run out of power, leading to the problem of energy

hole around the sink. In this case, even if the rest of nodes, specially the border ones, still hold sufficient energy, communication with the sink would be cut off, causing probably the end of the network lifespan.

5.3.2.1 The twofold compression pattern

To alleviate the overwhelming issue of energy hole, nodes' residual energies should be considered when selecting the set of representative nodes $\mathcal{N}_{rep}$. When all the nodes are directly connected to the sink, as in the star network topology, performing the selection cost function (5.3) is effective enough to attain the purpose of this work. Yet, when the data have to be forwarded via relaying nodes to reach the destination, taking into account only the transmitting source node residual energy is completely insufficient. Instead, the residual energy level of all the relaying nodes that would participate in the data forwarding should be assessed. Indeed, the metric (5.3) does not consider the entire route. Using (5.3), we will select sensor nodes with the highest residual energy, while ignoring the continuity ability of the entire route. Suppose a sensor node g^*, holding the maximum value of the combined metric (i.e. correlation-energy), is selected and there is a relaying node with a used up battery in its route towards the sink. In this case, the path will be cut off announcing probably the end of the network lifetime. Therefore, in addition to the correlation, a node is chosen to be a representative one under the condition that there is no depleted relaying node in its route. That is, the metric (5.5) is chosen for our EAMC for this scenario:

$$g^* = arg \max_{g \in S_1^j} \left(m_g'^2 \times \frac{E_g \times \min_{hp_g \in \mathcal{HP}_g} \left(E_{hp_g} \right)}{(\sum_{i \in \mathcal{N}_f} E_i)^2} \right), \tag{5.5}$$

where $\mathcal{HP}_g$ represents the set of nodes composing the route of the representative node g towards the sink[1]. In (5.5), adding the term $(\min_{hp_g \in \mathcal{HP}_g} \left(E_{hp_g} \right))$ means that we take into account also the relaying node with the lowest residual energy in the representative node selection process in order to avoid the fast depletion of the routes and hence the network partition, while there are still nodes with sufficient remaining energy that can forward data. Here, if the energy level of the relaying node hp_g that is belonging to the route of node g towards the sink is very low compared to other

[1]Note that $\mathcal{HP}_g$ contains only the relaying nodes and neither the representative node g nor the sink belongs to it.

nodes, the combined entity value will be weakened, and the node g won't be chosen as a representative node for the current detection period. As we can notice, in this energy-aware cost function, we have strengthened the weight of the factor $m_g^{'}$, which reflects how much the sensor g can represent the network, in order to maintain a good/efficient recovery quality. It will be shown in the simulations section that the introduced cost function is able to achieve an interesting and satisfactory trade-off between the data recovery quality and the network lifespan.

5.3.2.2 The single-level compression pattern

Generally, the multi-hop transmission is essential for the dense WSNs as well as for the the case of large networks (in terms of geographic distance), without being too much dense, where sensor nodes are far away from the sink. Particularly, in this kind of network, there is no need to make a significant number of sensor nodes completely inactive, for the entire current detection period, when executing data sensing. Accordingly, in this part, we won't pass through the selection of a set of representative sensor nodes. Instead, we proceed directly for the transmitting source nodes schedule. Furthermore, we want to evaluate our approach under the ordinary data sampling scenario, as well, in order to provide an overall work, where nodes can participate at least once during one detection period T. To do so, in each time slot t, using the same shared percentage pct_m, m_j transmitting source nodes are directly selected from the set CL_j of nodes composing the cluster j, according to (5.5), to sense the field and transmit their data readings to the sink. That is, instead of (4.13) and (4.14), we have:

$$m = \sum_{j=1}^{J} m_j, \tag{5.6}$$

where

$$m_j = pct_m\% \times cl_j. \tag{5.7}$$

Here, we proceed as if we set $pct_{Nrep} = 100$ and all the nodes are representative for the network. Certainly, there will be more computation than the twofold scenario, where the active node selection process, via (5.5), is effectuated only once for the entire detection period T. Fortunately, the one that is responsible for all that calculation is the sink, which is free of energy constraint. In fact, we assume that the sink

node has all the information regarding the sensor nodes' locations. Thus, it can compute, in advance, the energy to be consumed by the nodes for data sensing and forwarding. Thereupon, it is able to schedule beforehand the participation of the nodes during the entire detection period. In order to not increase again the communication overhead, the sink informs the concerned nodes about their data sensing schedule at the beginning of the detection period, i.e. we designate only one-shot scheduling transmission for the entire detection period.

Since, in each time slot t, energy consumption is uneven between nodes due to the multi-hop systems configuration, over a period of time we outface some sensor nodes whose routes hold relaying nodes with low residual energy. Performing (5.5) will keep these nodes out of the selection range for several successive time slots, until other nodes take their places. The fact of not being selected as a transmitting source node for successive time slots and not reporting data to the sink leads to the existence of successive missing entries in the received data matrix M. This sequence of missing data entries that may exist in the rows, referred to as a row structure fault in [24], impedes the MC resolution and highly increases the data reconstruction error. Therefore, for this single-level compression scenario, an extra step is added to the three-stage MC-based reconstruction pattern and set at the beginning of the recovery process, in order to detect the rows that hold structure faults and consider them as completely empty rows. This step consists simply in finding the sequence of successive zero entries holding a length larger than a given fixed size, which represents the minimum size of successive data missing from which that sequence is considered as a structure fault. That is:

$$StrFault_{min} = pct_{strF}\% \times T, \tag{5.8}$$

where, in accordance with the duration T of the detection period, pct_{strF} represents the parameter that fixes the minimum size of successive data missing from which the detected sequence is considered as a structure fault. It will be shown in the simulation part that treating separately the rows that hold structure faults significantly improves and refines the data reconstruction accuracy.

5.4 Numerical Resutls

In chapter 4, we have compared the performance of the proposed OCBMC versus the scheme of [24] that had treated a relatively similar scenario to our twofold data loss one. We have found that our structured approach outperforms the baseline scheme in terms of both data reconstruction error and overall network energy consumption. For that reason, in this chapter, we have based on this comparison to carry on with our structured scheme and improve its design and techniques. The proposed energy-aware data gathering EAMC, where energy is jointly taken into account with the correlation criteria, is compared to the OCBMC scheme. This simulation will reveal the impact of the updated selection cost function on the trade-off between the data recovery accuracy and the network lifetime under all the investigated scenarios, and for both types of sensor nodes. Then, to summarize and confirm our results, this trade-off is evaluated in a different manner.
Thereupon, in order to estimate the data reconstruction accuracy for the implemented schemes, we opted for the metrics (4.17) and (4.20) of chapter 4. To simulate the implemented schemes and evaluate their performance under different CRs, we vary pct_{Nrep} from 10 to 60, and for each given pct_{Nrep}, we vary pct_m from 10 to 80. Regarding the network parameters, we consider that $N = 50$ sensor nodes are randomly deployed in a square observation area of size $100m \times 100m$, and we monitor the WSN throughout a detection period of length $T = 100$ time slots.

In these simulations, we focus on the principal purpose of this work, which is the network lifetime improvement. For that reason, we analyze the performance of the EAMC, where we consider for each scenario and network topology the adequate cost function. Namely, we evaluate the metric (5.3) for the single-hop star topology and the metric (5.5) for the multi-hop mesh topology. Moreover, we assess the proposed approach under both of sensor nodes types; the ordinary sensors, where the energy consumed in data sensing is quite low, and the specific power greedy ones, where the acquisition energy cost is greater than that of the communication cost [9]. The parameters of the used energy consumption model are outlined in the Table 5.1.

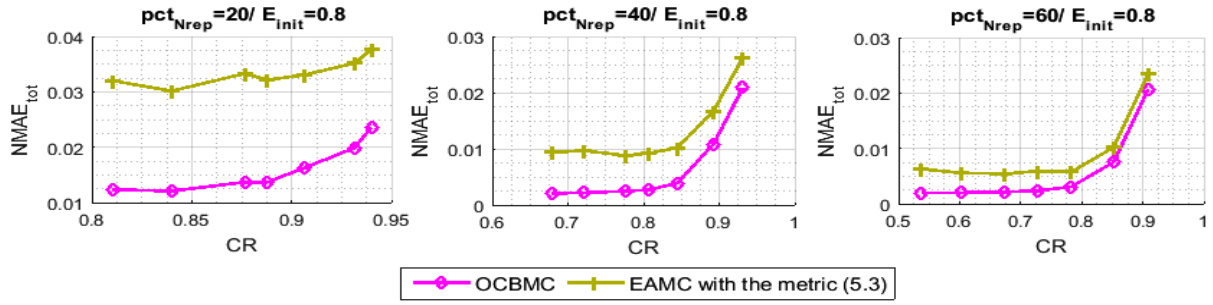
To begin, we consider the single-hop star network and we compare the EAMC approach to the OCBMC. Figures 5.1 and 5.2 depict the trade-off between the $NMAE_{tot}$ and the network lifetime. The network lifetime denotes the number of detection period

Table 5.1: Simulation Parameters for energy consumption

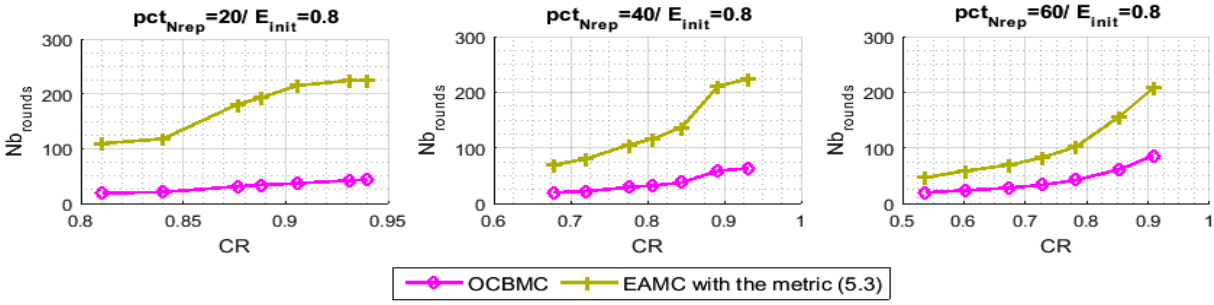
Parameter	Ordinary sensor node	Greedy power sensor node
E_{init}	0.8 J	40 J
I_{sens}	50 μA [91]	25 mA [8]
T_{sens}	0.5 mS [8]	0.5 mS [8]
V_{sup}	2.25 V [91]	2.7 V [8]
$E_{elec-tr}$	50 nJ/bit [92]	
$E_{elec-rc}$	5 nJ/bit [92]	
ε_{amp}	100 $pJ/bit/m^2$ [8]	
L	1024 $bits$	

T that a scheme can achieve without causing the death of any sensor node in the network, i.e. Nb_{rounds}. Indeed, the first node that exhausts all its battery energy announces the death of the network and determines its lifetime Nb_{rounds}. Note that for each case, when we vary the compression ratios pct_{Nrep} and pct_m, the $NMAE_{tot}$ and the Nb_{rounds} are simultaneously calculated then depicted in Figures 5.1 and 5.2 for both compared approaches. Moreover, the final depicted $NMAE_{tot}$ represents the average of all the resulting $NMAE_{tot}$ during the ensured Nb_{rounds}[2]. Integrating the residual energy with the correlation, as a second weighty factor, will certainly lighten the impact of the correlation on the data recovery quality. However, in this chapter, we target to reach a robust and equitable compromise between the two addressed factors. As we can note, we still achieve a sufficiently good data recovery accuracy even for small values of pct_{Nrep} (i.e. when there is a significant number of completely empty data rows in M), while at the same time the network lifetime is highly improved. As an example, with the ordinary sensor nodes ($pct_{Nrep} = 20, pct_m = 10$), the $NMAE_{tot}$ passes from 0.023 to 0.037 with the EAMC, while the network lifetime is expanded with a percentage of 416.94% (i.e. Nb_{rounds} passes from 43.34 to 224.04 rounds). On the other hand, with the greedy power sensor nodes ($pct_{Nrep} = 20, pct_m = 10$), the $NMAE_{tot}$ passes from 0.029 to 0.038 with the EAMC, while the network lifetime is expanded with a percentage of 303.2% (i.e. Nb_{rounds} passes from 22.52 to 90.8 rounds). Moreover, we can notice that as the number of active nodes is increased, the gap of $NMAE_{tot}$ between the two compared algorithms is significantly reduced,

[2]The $NMAE_{tot}$ and the Nb_{rounds} of the simulations of Figures 5.3, 5.4, 5.5, 5.6 and 5.7 have been calculated following the same manner.



(a) The $NMAE_{tot}$ for OCBMC and EAMC approaches.

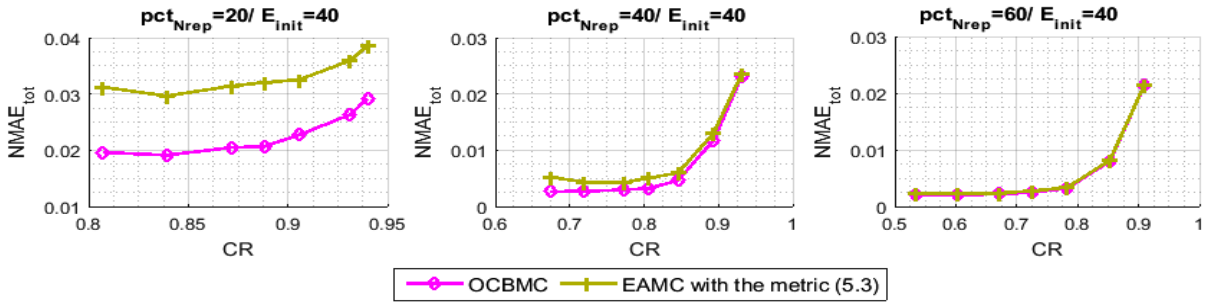


(b) Nb_{rounds} for OCBMC and EAMC approaches.

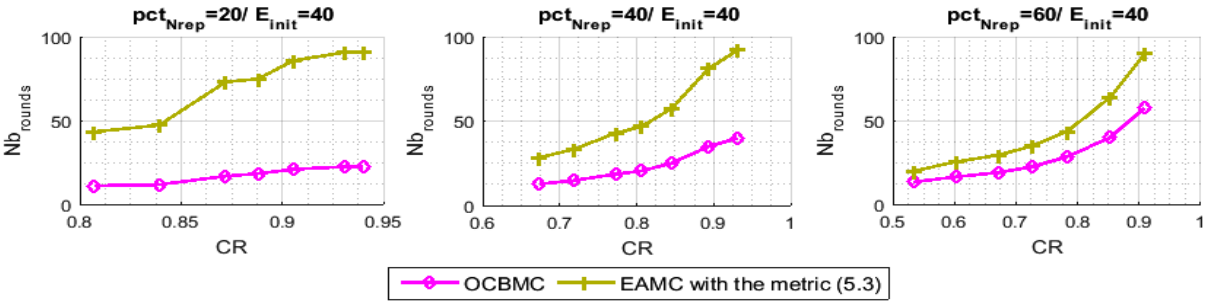
Fig. 5.1. Performance trade-off between the data reconstruction error and the network lifetime for OCBMC and EAMC approaches in the single-hop star topology with ordinary sensors.

whereas, that of the network lifetime is still clearly noteworthy. Precisely, with the greedy power sensor nodes, starting from $pct_{Nrep} = 40$, we start to reach gains on the network lifetime almost without deteriorating the $NMAE_{tot}$.

Let us now focus on the second scenario: the twofold data compression in the multi-hop mesh network topology. For the ordinary sensor nodes, where the consumed energy during data detection is quite low compared to that used for data transmission, the trade-off between the $NMAE_{tot}$ and the Nb_{rounds} is illustrated in Figure 5.3. Particularly, as we can see, we keep intentionally considering the performance comparison of the $NMAE_{tot}$ that has been performed in the simulation of Figure 4.13 between the CBMC and the OCBMC. Interestingly, it is noteworthy that even though the $NMAE_{tot}$ is slightly increased when considering the sensor residual energy in the metrics (5.3) and (5.5), it is still quite inferior to that given by the original CBMC,



(a) The $NMAE_{tot}$ for OCBMC and EAMC approaches.



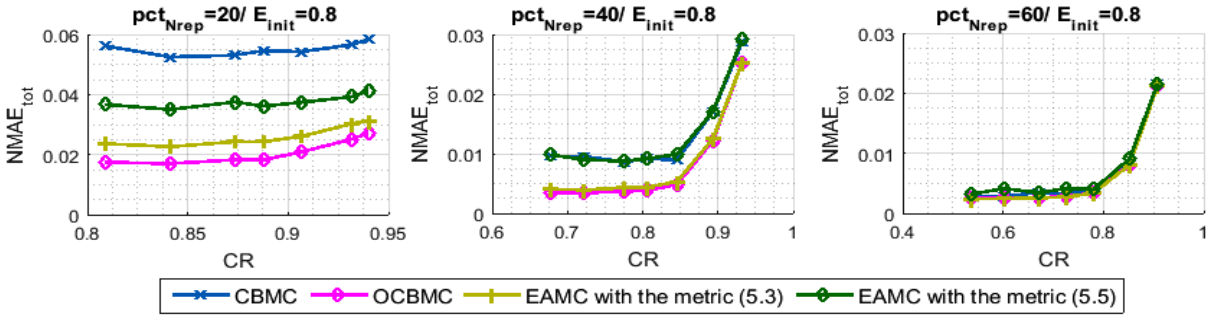
(b) Nb_{rounds} for OCBMC and EAMC approaches.

Fig. 5.2. Performance trade-off between the data reconstruction error and the network lifetime for OCBMC and EAMC approaches in the single-hop star topology with the greedy power sensors.

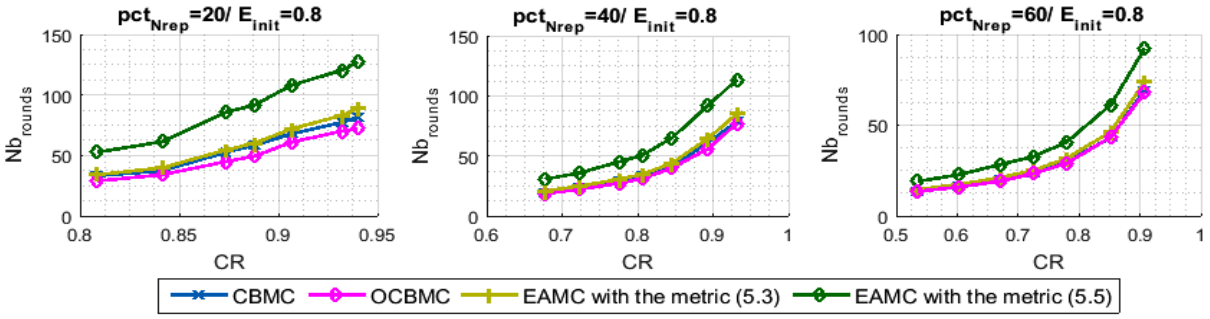
when the correlation criteria is not taken into account. Let us now compare the metrics (5.3) and (5.5), we can observe that the metric (5.5) achieves a better Nb_{rounds} at the cost of a slight increase of the $NMAE_{tot}$. Indeed, since the entire route is considered with (5.5), sensor nodes having depleted relaying nodes in their paths towards the sink are less susceptible to be selected as representative nodes. Here, it is worth mentioning that as long as we keep achieving a sufficiently good recovery quality (i.e. a low $NMAE_{tot}$), we privilege the second crucial factor that is the network lifetime expanding. As we can see in Figure 5.3, for pct_{Nrep} equals to 60, the resulting data recovery error, when we perform the metric (5.5), is almost the same as when we use (5.3), whereas, the Nb_{rounds} ensured by the retained metric (5.5) is higher than that given by (5.3). For example for $(pct_{Nrep} = 60, pct_m = 10)$, with the same $NMAE_{tot}$, performing the EAMC using the cost function (5.5) can prolong the network lifetime with a percentage of 35.69% compared to the OCBMC, whereas, the yield of (5.3) is limited to 8.7%. This is because metric (5.5) takes into account the entire route through the value of $(\min_{hp_g \in \mathcal{HP}_g} (E_{hp_g}))$. Hence, this technique is able to ensure a much longer lifetime for the network, when the sensor nodes are ordinary ones. Another example, for $pct_{Nrep} = 40$ and $pct_m = 10$, by increasing the $NMAE_{tot}$ from 0.025 to only 0.029, the metric (5.5) can prolong the network lifetime with a percentage of 47.28%.

Nevertheless, when the deployed sensor nodes are greedy power ones, as it has been depicted in Figure 5.4, the performance of both metrics (5.3) and (5.5) become very close. Indeed, with this type of sensor nodes, the amount of energy that is consumed in data forwarding by the relaying nodes becomes much less than that consumed in sensing by the assigned transmitting source nodes. Consequently, both metrics tend to choose the same set of representative nodes since the amount E_g (i.e. the residual energy of the representative node of interest g) represents, with this type of nodes, the most important and the dominant component that heads the active node selection process. We can distinctly note the very significant improvement brought by the EAMC, with both metrics (5.3) and (5.5), compared to the OCBMC in terms of network lifetime, especially for high CRs. As an example, for $(pct_{Nrep} = 40, pct_m = 10)$, we can reach an amelioration of 124.4% in terms of Nb_{rounds} with both metrics, while still maintaining nearly the same $NMAE_{tot}$ compared to the OCBMC.

In order evaluate the scalability of the proposed solution, we have compared in Fig-

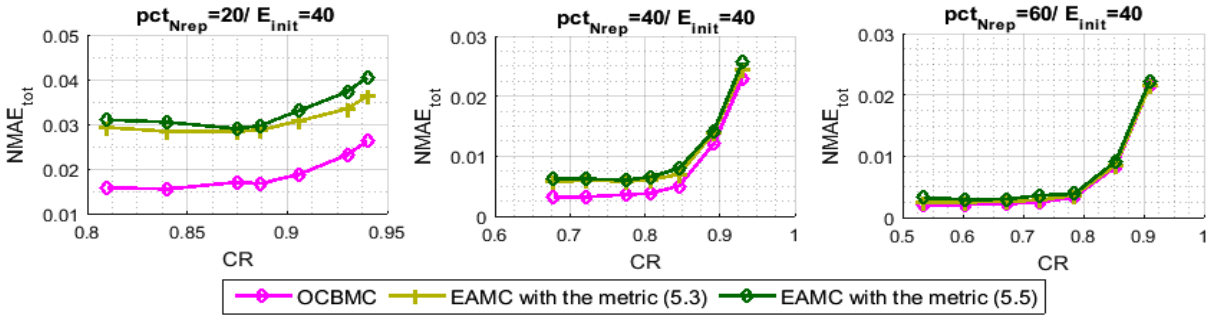


(a) The $NMAE_{tot}$ for CBMC, OCBMC and EAMC.

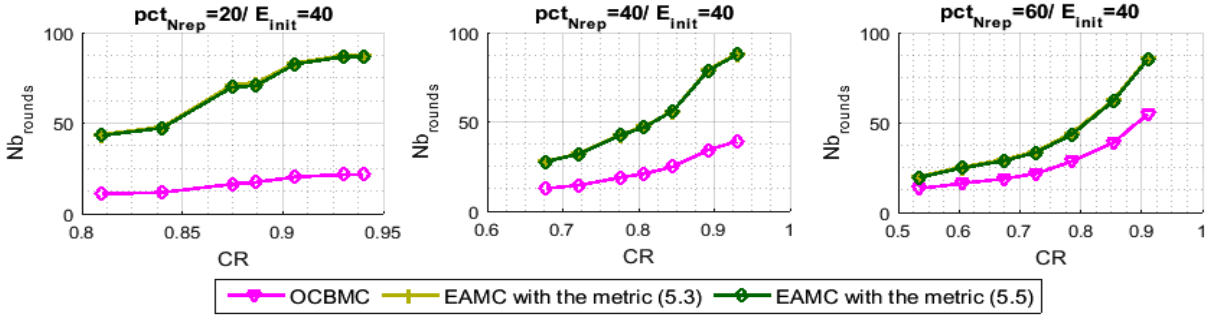


(b) Nb_{rounds} for CBMC, OCBMC and EAMC.

Fig. 5.3. Performance trade-off between the data reconstruction error and the network lifetime for the compared approaches in the twofold compression scenario and multi-hop mesh topology with ordinary sensors.



(a) The $NMAE_{tot}$ for OCBMC and EAMC approaches.



(b) Nb_{rounds} for OCBMC and EAMC approaches.

Fig. 5.4. Performance trade-off between the data reconstruction error and the network lifetime for the compared approaches in the twofold compression scenario and multi-hop mesh topology with greedy power sensors.

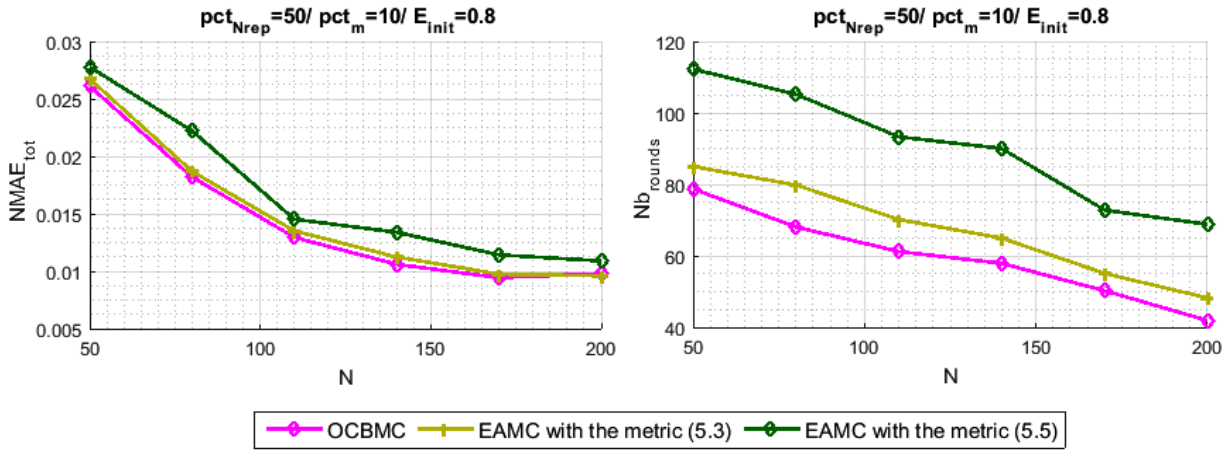


Fig. 5.5. Performance trade-off between the data reconstruction error and the network lifetime for the compared approaches in the twofold compression scenario and multi-hop mesh topology with ordinary sensors and with respect to the number of sensor nodes.

ure 5.5 the performance of the implemented schemes, for the scenario of Figure 5.3, with respect to the number of sensor nodes. To this end, for $pct_{Nrep} = 50$ and $pct_m = 10$, we vary the number of the deployed sensor nodes, and for each case we measure the $NMAE_{tot}$ and the Nb_{rounds}. As we can note from this simulation, the EAMC with the metric (5.5) keeps outperforming both the EAMC with the metric (5.3) and the OCBMC in terms of Nb_{rounds} with respect to N, although it increases a little bit the $NMAE_{tot}$. These simulations confirm the scalability of the proposed scheme and show that it keeps the same behaviors, even when we increase or decrease the number of sensor nodes N. Note that the data recovery error is reduced, for all the compared schemes as N is raised since the MC-based reconstruction methods work very well for large-scale data estimation problems. Regarding the network lifetime, we can note that it is reduced as N is raised since the overall energy consumption is increased due to packet relaying.

In scenario three (i.e. the single-level compression scenario in the multi-hop mesh topology), whose results are depicted in Figures 5.6 and 5.7, we have compared the EAMC scheme using the metric (5.5) with its original version, the CBMC, for $pct_{Nrep} = 100$. Indeed, since in the WSNs, most of the time, the signal is time-stationary, using only the correlation criteria via the OCBMC to seek for the m

transmitting source nodes in each time slot t leads to probably having the same transmitting source nodes during all the detection period T. The resulting configuration entails a schedule, where the same chosen source nodes will transmit their data during all the time slots t composing the detection period T, while the rest of nodes remain completely inactive. The OCBMC is not suitable for this scenario since, in this part, we aim to address an ordinary data sampling scenario, where nodes can participate in data sensing and transmission at least once during one detection period T. As we can note from Figure 5.6, the $NMAE_{tot}$ achieved by the EAMC (i.e. the dark green curve) gets worse compared to the original CBMC, despite the amelioration achieved in terms of Nb_{rounds}. This is due to the existence of the row structure faults that appear in the data vectors corresponding to the nodes that are remaining outside the range of selection for several successive time slots. In addition to the fully empty data rows, the structure faults are among the serious obstacles that not only impede the MC resolution but also pollute the received data [24]. For that reason, before applying the MC method, the rows holding these structure faults should be removed from M then recovered through stage two if the node corresponding to this row is an absent node or recovered through stage three if the node is an Isolated one. In [24, Section. V], authors have proposed an algorithm that detects rows holding structure faults. This technique is implemented here and the resulting performance are depicted with the dotted black curve. As we can note from the figure legend, this technique is dependent to two different parameters $N^{'}$ and α. Altogether, these parameters give $\chi^2_{N',\alpha}$, which represents the upper α percentage point of the chi-square distribution with the degree of freedom $N^{'}$. Although they had shown how to choose $N^{'}$, the selection of α has been done without any explanation, and according to our several simulations, it should be noted that the slightest variation of any of these parameters makes an important difference in the $NMAE_{tot}$ performance. Here, the value of α has been determined empirically[3]. As for our proposed structure faults detector (5.8), it has been evaluated with respect to a threshold parameter pct_{strF} that, in accordance with the detection period duration T, fixes the minimum size of successive missing entries from which the sequence is considered as a structural fault. Surprisingly, we can clearly note from Figure 5.6 that, despite its simplicity, our proposed method can significantly reduce the data recovery error of our EAMC for the entire range of CR, while still keeping the same Nb_{rounds}, whereas, the improvement brought by the technique

[3]We report here the simulation of only the most performing value.

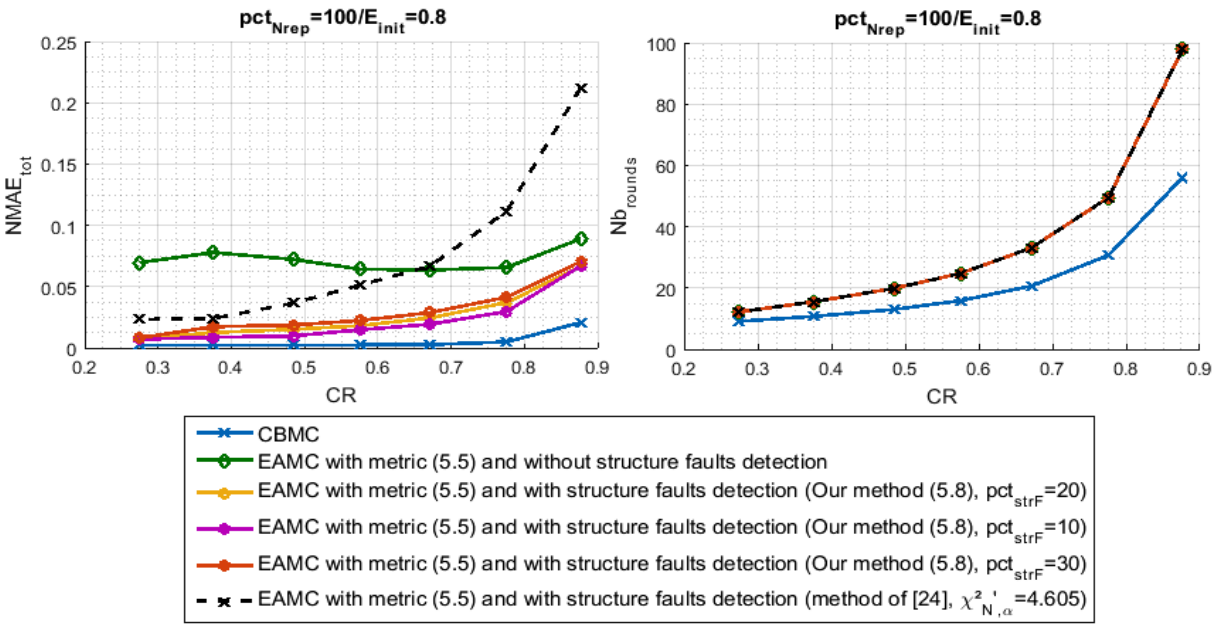


Fig. 5.6. Performance trade-off between the data reconstruction error and the network lifetime for the compared approaches in the single-level compression scenario and multi-hop mesh topology with the ordinary sensors.

of [24] is limited by a restricted range of CR. Moreover, the $NMAE_{tot}$ given by the EAMC with our structure fault detection method is not only lower than that given by the EAMC without structure fault detection, but also it is lower than the $NMAE_{tot}$ resulting from the EAMC with the method of [24], which unfortunately makes the data recovery accuracy worse for the high CRs (i.e. $CR > 0.66$). In fact, using the same parameter value α for both high and low CRs impedes the imposed threshold for the structure faults detection from maintaining a low error ratio across the entire range of CR. These simulations show that our technique is not only simpler than that proposed in [24], but also it is more efficient.

Figure 5.7 depicts the simulations obtained with the greedy power sensors. It shows a significant decrease in the $NMAE_{tot}$ corresponding to EAMC compared to those of the ordinary low power sensors of Figure 5.6. Since the amount of consumed energy in data forwarding by the relaying nodes becomes far less than that consumed in sensing, we do not have paths that run out quickly, and the choice of the active transmitting source node becomes directed only by the energy of the node in question, not the

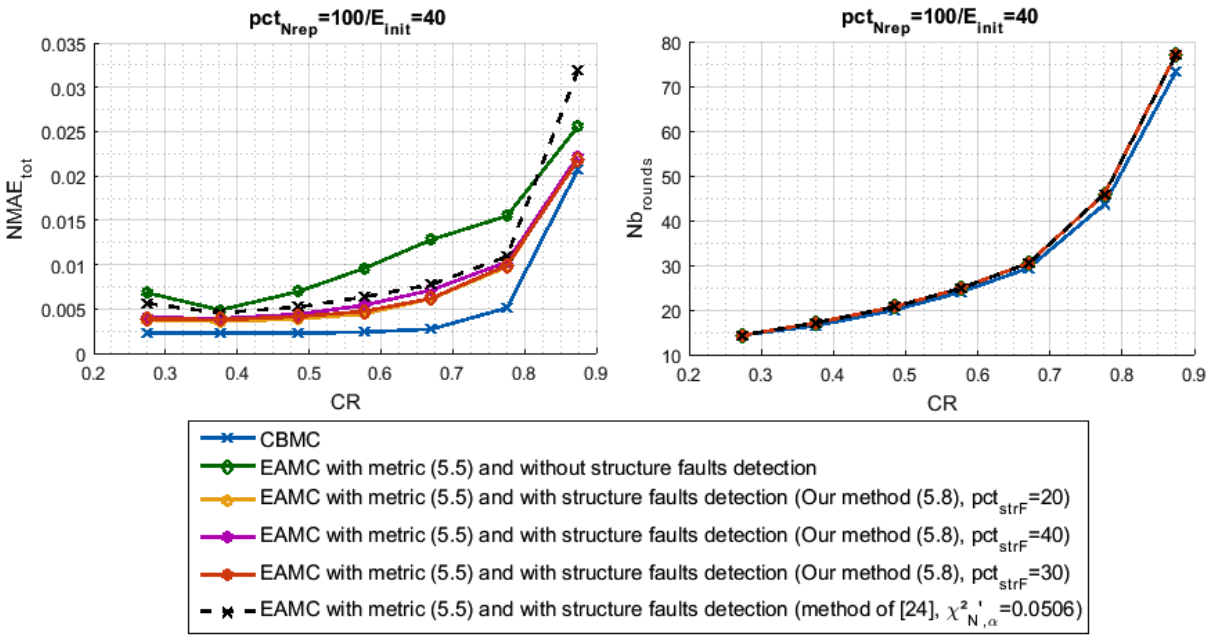


Fig. 5.7. Performance trade-off between the data reconstruction error and the network lifetime for the compared approaches in the single-level compression scenario and multi-hop mesh topology with the greedy power sensors.

relaying nodes composing its path towards the sink. As a result, the number of the structural faults as well as their sizes are reduced. For that reason, we vary again the parameter α in this simulation since the one used in Figure 5.6 was not suitable for the present case. As for the network lifetime performance, unexpectedly, for this kind of data compression scenario and when the deployed sensors are greedy power ones, the energy consumption is too great that we can make significant improvements in terms of Nb_{rounds}.

In the final part of the simulations section, the trade-off between the data recovery error $NMAE_{tot}$ and the network lifetime, measured with Nb_{rounds}, has been investigated in a more realistic manner. For a given $pct_{Nrep} = 30$, an error ratio upper bound is fixed. Here, the *Fixed* $NMAE_{tot}$ is varied from 0.02 to 0.1, and for each value we compute the maximum number of detection periods that the scheme can ensure, despite the eventual existence of dead sensor nodes. As long as the implemented scheme can achieve an $NMAE_{tot}$ lower or equal to the fixed bound, the network is considered as operational. When the data recovery error ratio of this scheme exceeds

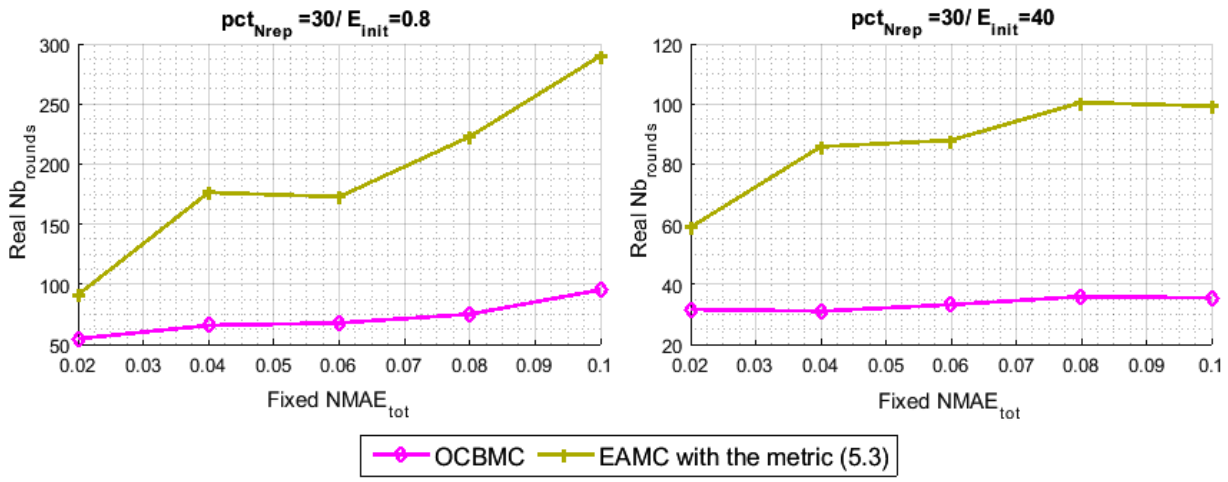


Fig. 5.8. Real network lifetime vs. fixed upper bound data recovery error ratio for the compared approaches in the single-hop star topology with both types of sensor nodes.

the $Fixed\ NMAE_{tot}$ value, we consider the network as dead. Doubtless, if a sensor node runs out of energy, it can no longer participate in data sensing and forwarding.

Likewise the previous simulations, we start with the single-hop star topology. We can note from Figure 5.8 that the EAMC, using the metric (5.3), can highly prolong the network lifetime compared to the OCBMC, while still able to maintain a data reconstruction quality better than the imposed level. Another interesting point is that the potential improvement of EAMC is higher than OBMC in terms of network lifetime. In other words, the $Real\ Nb_{rounds}$ achieved by the EAMC increases faster than that achieved by the OCBMC, when the upper bound error ratio is expanded.

Figure 5.9 depicts the obtained results for scenario two, which is the twofold data compression in the multi-hop mesh network topology. As we can notice, the obtained performance results confirm those of Figure 5.3 and Figure 5.4 and prove the efficiency of the proposed EAMC with both types of sensor nodes. Particularly, we can perceive the existence of drop points in the plots, namely, $Fixed\ NMAE_{tot} = 0.06$ with the ordinary sensor nodes and $Fixed\ NMAE_{tot} = 0.1$ with the greedy power sensor nodes. Since we perform a discretized CR to reach an $NMAE_{tot}$ that is less than or equal to the fixed one, i.e. pct_m varies from 10 to 70 with step of 10, there is possibility that a CR under the needed one is used (pct_m above the needed one), and some

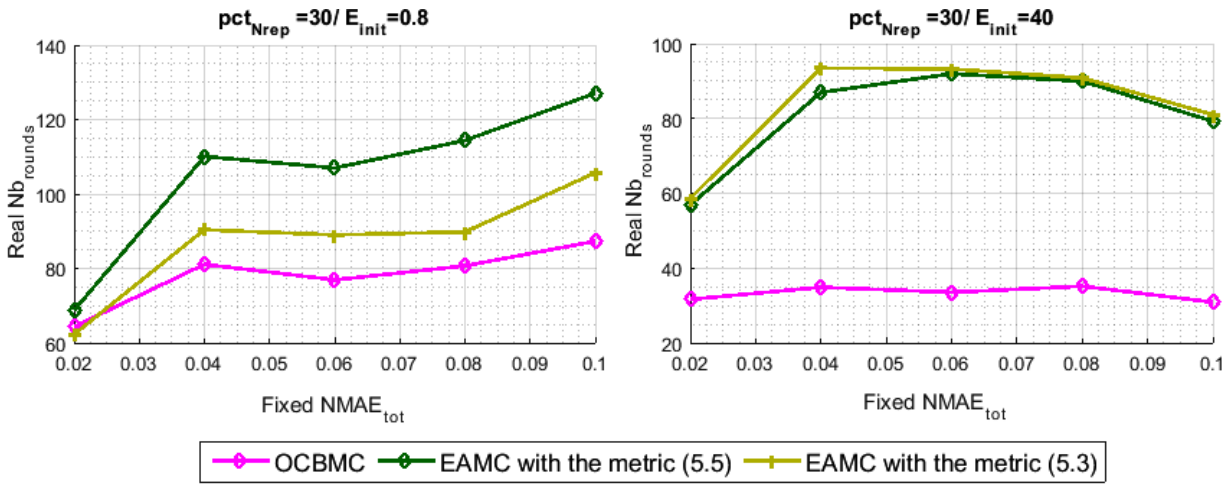


Fig. 5.9. Real network lifetime vs. fixed upper bound data recovery error ratio for the compared approaches in twofold compression scenario and multi-hop mesh topology with both types of sensor nodes.

specific nodes may untimely die. Consequently, the achieved $NMAE_{tot}$ increases and exceeds the upper bound $Fixed\ NMAE_{tot}$, which causes the death of the network. The latter explanation denotes that a piecewise effect appears in these points. As we can note, the appearance of these particular points is clearer with the case of the greedy power sensor nodes, where the data sensing activities carried out by the transmitting source nodes consume the bulk of the total power consumption, and the selection of the active nodes is mainly based on the residual energy of the sensor of interest. Nevertheless, note that what mostly interests us is the difference, in terms of network lifetime performance, between the curves that still exists even with the drop points.

5.5 Conclusion

In light of the importance of the energy saving and lifetime for wireless sensor nodes that are suffering from a limited power capacity, in this chapter, we have presented an adaptive data collecting scheme, called the EAMC. Since the all-node-active condition is completely impractical for WSNs with the energy constraints, the proposed approach can dynamically designate the transmitting source sensor nodes that can

afford the sustainability as long as possible of the network lifetime and alleviate the problem of energy load imbalacing, according to an energy-aware cost selection function. Additionally, the proposed EAMC scheme aims not only for achieving the energy efficiency for the network but also for preserving a sufficiently good quality of data reconstruction as it still takes into account the correlation criteria among sensors in order to select those who can best represent the network. Furthermore, we have evaluated our approach under different network topologies and scenarios, while performing, in each time, the adequate energy-aware metric. Moreover, to recover the entire data matrix, despite the existence of a significant number of completely missing rows corresponding to the inactive nodes, we have relied on the three-stage data reconstruction framework of the previous chapter. For the last addressed scenario, to refine and enhance the data recovery quality, we have added a simple step to the data recovery techniques, which efficiently detect the structure faults that may appear in the received data matrix. Simulations have proven that the proposed scheme can achieve an interesting trade-off between the data reconstruction accuracy and the network lifetime compared to the baseline schemes. Accordingly, the EAMC scheme can be considered as very interesting for research in the field of energy saving since, particularly, it is able to efficiently overcome the twofold data loss scenarios.

Chapter 6

Conclusion and Perspectives

6.1 Summary of Contributions

In WSN, sensors are deployed in order to collect periodically measures of physical fields, which can be related to a wide range of applications as security, science, industry, civil infrastructure, etc. However, their crucial nature of limited power, memory and computational capacities requires focusing on minimizing energy consumption and processing complexity to ensure a longer lifetime for the network [72]. Thereupon, the purpose of this thesis is to establish and evaluate energy-efficient data gathering schemes for future WSNs. Relying on the CS and the MC methodologies, this dissertation proposes three different approaches.

We started by evaluating an uncommon space-time CS-based design, where we have performed a strategy that neatly and determinately selects a subset of active sensor nodes under the criterion of having the best presentation of the whole network, using a correlation-based metric, when, at the same time, they are "near" the sink. These designated nodes will deliver their data readings to the sink for only the same subset of predetermined time slots, i.e. the data signal is adaptively sub-sampled in the space as well as temporal dimension. This is different to the existing spatial CS data gathering patterns, where, in each time slot, the sensors' readings are linearly combined along a multi-hop routing. Surprisingly, recovering the entire data with such unfamiliar or unusual situation has worked successfully thanks to the use of an adaptive spatial sparsifying basis Ψ_S with a covariogram-based estimation. Since the addressed data

gathering strategy reduces dimensionally the number of data samples in space and time, the Kronecker framework has been performed in the data reconstruction process for the proposed STCS-RA approach to take advantage of the signal sparsity in both dimensions.

The remaining schemes of this dissertation focused on the application of the MC methodology because of its numerous benefits. In the second contribution, we have developed a structured MC-based framework that is able to deal with the existence of a significant number of completely missing data rows in the received data matrix. These empty rows result from the inactive nodes that do not participate at all in the data sensing process during the entire detection period. The reconstruction of the entire data has been achieved successfully with high accuracy thanks to the proposed minimization-based interpolation technique, which is annexed, as a third stage, to the MC resolution. Furthermore, since we are mainly interested in the high data loss scenarios, gathering the limited amount of data to be transmitted from the active nodes must be neatly scheduled to afford the sufficient information about the whole network area. For that reason, we have proposed the CBMC and the OCBMC data gathering approaches, which assign the sensor nodes into groups using a data-based spectral clustering technique. The detected clusters are taken into account in the representative nodes (active nodes) selection process then in the data sensing schedule with the use of the same shared percentage between clusters in order to provide an equitable representation of the monitored area. Through simulations, we have shown that such an adaptive data sampling deeply affects the recovery quality of not only the missing data corresponding to the active nodes but also those corresponding to the completely inactive ones.

Aiming to further optimize the use of WSNs resources, we present in our third contribution an adaptive EAMC data gathering approach that extends the introduced scheme of the second contribution. The proposed data gathering strategy has been conceived with the intention of systematically maintaining a load balancing among nodes and maximizing the network lifetime, while still achieving a low data reconstruction error. Indeed, in addition to the correlation, we have incorporated the sensors' residual energies in the representative node selection process and developed different combined energy-aware cost selection metrics. Depending on the variation that occurs on the nodes inter-correlation as well as on their available power supplies,

the proposed approach selects nodes that can best represent the network, taking into account the efficiency of the network energy utilization. We have evaluated our approach under different network topologies and scenarios, while performing, in each time, the adequate energy-aware metric. For each case, the trade-off between the data recovery error and the network lifetime is measured, and the performance behaviour of the proposed data gathering approach is studied for both types of sensor nodes; the low-power sensors and the hungry-power ones.

6.2 Perspectives

The solutions suggested throughout this dissertation permit the rise of some new insights and ideas, which can further ameliorate the WSNs performance. Indeed, a number of mechanisms developed and proposed in this work can be extended and updated, and then performed in another domain or in a variety of manners.

6.2.1 STCS iterative reconstruction using an adaptive Ψ_T

Although spatial and temporal correlations have been jointly exploited with the STCS, as both distributed and local CS have been applied for data compression, and the Kronecker CS framework have been performed for decompression, only the spatial matrices Φ_S and Ψ_S have been adaptively computed according to the signal variation. In this context, a natural modification is to estimate a temporal sparsifying basis Ψ_T that can be dynamically adapted to the time-varying statistics of the signal field. Thus, as a perspective, we propose an iterative algorithm, where the estimation of X_{cs} can be progressively refined. Suppose that $\widehat{X}_{cs}$ is reconstructed based on the received measurements, as described in chapter 3. Then, using this $\widehat{X}_{cs}$ as a first estimation denoted by $\widehat{X}_{cs_1}$, we can improve the accuracy of $\widehat{X}_{cs_2}$, of the same current sensing period, after recalculating Ψ_T through replacing DCT by the PCA basis or another data dependent and advanced temporal sparsifying basis in order to better exploit the intra-sensor correlation. Note that we can refer to a new iteration Itr each time $\widehat{X}_{cs_{Itr}}$ is re-estimated using $\widehat{X}_{cs_{(Itr-1)}}$. Besides, for the next sensing periods, these estimations will be used and the precision may increase. Here, for example, an

experimental study can be done to fix the number of re-estimations (iterations) that one should perform to reach a required error ratio.

6.2.2 From a centralized approach to a distributed one

In the proposed approaches, a centralized node, which is the the sink, is the one that is responsible for the selection of the representative nodes and for their data sensing activities schedule over each detection period. Even though, this meets well the constrained resources and computational capacities of the deployed wireless devices, it may be more desirable to distribute the computation of the active node selection algorithm and the data sensing schedule between nodes in order to make them more autonomous. Moreover, suppose that the sink has a finite power supply, as in many practical applications. Thus, establishing an adaptive data gathering scheme, with a decentralized manner, can significantly reduce the computational complexity carried by the sink node and even speed the data gathering process. To this end, performing this purpose, while keeping in mind the overall network energy capacity and efficiency, makes it an extremely challenging and worth pursuing research issue.

6.2.3 The three-stage MC-based reconstruction approach in Massive MiMo

In massive Multi-Input Multi-Output (MIMO) systems, a precise acquisition of the Channel State Information (CSI) is needed for signal detection, resource allocation, beamforming, etc [93]. Yet, with the explosive growth of the single-antenna user terminals number, the Base Station (BS) should estimate channels that are associated with hundreds of users, leading to high pilot overhead. The idea here is to let only a small number of users transmit their pilots in the training phase of each coherence interval and, using the proposed three-stage MC-based reconstruction approach of chapter 4, the BS will estimate all channels, even those corresponding to users who have not sent pilot signals. This framework can be implemented and introduced as a channel estimation scheme for the uplink massive MIMO systems based on the assumption of channel reciprocity in the Time Division Duplexing (TDD) mode [94]. Since most of wireless channels are sparse, the MC method can represent a suitable

solution for channels estimation [95], [96]. To model the considered massive MIMO system, we assume that the BS is equipped with an array of a significant number M of antennas. For the uplink mode, to estimate the channel matrix $H \in \mathbb{C}^{M\times K}$, the BS receives training signal vectors of pilots $\Phi = [\phi(1)^{tr}, \phi(2)^{tr}, ..., \phi(K)^{tr}]^{tr}$ sent by a large number K of users, where $K \leq M$. Conventionally, for each coherence interval, each user should transmit a pilot sequence of length L in the training phase, where $L \geq K$. Accordingly, Φ represents the $K \times L$ total training matrix that consists of K L-length training pilot sequences, and the received signal matrix, denoted by $Y \in \mathbb{C}^{M\times L}$, is given by the following equation [97]:

$$Y = H\Phi + N. \tag{6.1}$$

In (6.1), $N \in \mathbb{C}^{M\times L}$ represents the additive white Gaussian noise matrix, whose entries are i.i.d $\mathcal{N}(0, \sigma_N^2)_{\mathbb{C}}$.
To reduce the number of transmissions during the coherence time, we assume that a small number $K_{rep} \ll K$ of the users will transmit their L-length training pilot sequences, i.e Φ_{rep} of size $K_{rep} \times L$, and instead of finding $H \in \mathbb{C}^{M\times K}$, the BS would firstly estimate a sub-matrix $H' \in \mathbb{C}^{M\times K_{rep}}$ using the MC method with its noisy version. As an example, paper [97] had provided a mathematical MC-based formulation of the problem (6.1) in Eq. 7 and developed a solution in Eq. 11. Here, for our case, we have to solve the equation (6.2) instead of (6.1) and find H':

$$Y' = H'\Phi_{rep} + N'. \tag{6.2}$$

Secondly, the BS updates $H' \in \mathbb{C}^{M\times K}$ by adding the $(K - K_{rep})$ empty columns, which correspond to the users that did not sent their training pilot sequences, and placing them in their proper locations of H. Finally, it carries on with stage 2 and stage 3 to estimate these remaining columns to get the entire $M \times K$ channel matrix H.

6.2.4 The EAMC data gathering scheme with a dynamic routing

Using the already established routes with a static routing protocol, in data forwarding, may limits the performance improvement that an energy-aware data gathering

scheme can achieve. An interesting practical consideration is to update the paths systematically according to the remaining energy levels of the relaying nodes in order to further prolong the network lifespan. Indeed, it is noteworthy that cross layer optimization may achieve a considerable performance improvement. Hence, our idea here is to to keep using the representative node selection cost function (5.3) even in the multi-hop mesh topologies in order to preserve a better data recovery accuracy. However, instead of forwarding data to the sink through static paths, for example, each sensor node would choose as its next hop the sensor node, within its range obviously, that has the highest residual energy and at the same time can achieve the largest geographical advancement toward the sink. To select the appropriate forwarder, the balance between the residual energy of the node of interest and its distance toward the sink can be modeled by a certain cost function. Doing that may further improve the network lifetime, Nb_{rounds}, while maintaining the same low $NMAE_{tot}$.

Appendices

Appendix A

Publications of the thesis

1. M. Kortas, O. Habachi, A. Bouallegue, V. Meghdadi, T. Ezzedine, and J. P. Cances:“ The Energy-Aware Matrix Completion based Data Gathering Scheme for Wireless Sensor Networks,“ in IEEE Access Journal, 2020
2. M. Kortas, O. Habachi, A. Bouallegue, V. Meghdadi, T. Ezzedine, and J. P. Cances: “ Energy efficient data gathering schema for wireless sensor network: A matrix completion based approach,“ in IEEE Software, Telecommunications and Computer Networks (SoftCOM), 2019
3. M. Kortas, V. Meghdadi, A. Bouallegue, T. Ezzedine, O. Habachi, and J. P. Cances: “Routing aware space-time compressive sensing for wireless sensor networks, “ in IEEE Personal, Indoor, and Mobile Radio Communications (PIMRC), 2017
4. M. Kortas, A. Bouallegue, T. Ezzedine, V. Meghdadi, O. Habachi, and J. P. Cances: “Compressive sensing and matrix completion in wireless sensor networks, “ in IEEE International Conference on Internet of Things, Embedded Systems and Communications (IINTEC), 2017

Appendix B

Extra Simulations

B.1 Spatial Correlation feature

To evaluate the spatial dependency between the deployed sensor nodes, we use a kind of an $N \times N$ binary symmetric matrix Y_c, called a 1-hop topology matrix, where both columns and rows denote the sensor nodes. We assign 1 to $Y_c(i,j)$ and $Y_c(j,i)$ if we finds that sensor node i and sensor node j are 1-hop neighbors. But, according to the signals nature that we consider, we assume that even though two sensor nodes are geographically close to each other, if they don't belong to the same portion field D_h, they are not considered as neighbors. Since spatial correlation is mostly apparent between nearby sensor nodes, we compute the normalized difference between the data reading $X(i,t)$ of node i in time slot t with the sum of data readings of its one-hop neighbors [24]. That is:

$$\Delta S_{gap}(i,t) = \frac{X(i,t) - (Y_{c_{(i)}} X^{(t)} / \sum Y_{c_{(i)}})}{mean_h(diff)}, \tag{B.1}$$

where $X^{(t)}$ is the t^{th} column of X, $Y_{c_{(i)}}$ is the i^{th} row of Y_c and $mean_h(diff)$ represents the average of the calculated differences between the maximum and minimum data samples $X(i,t)$ discretized from the H fields' portions, i.e. the average of the largest differences between data samples of the H fields' portions. Figure B.1 plots the cumulative distribution function (CDF) of the ΔS_{gap} values. We can note that the probability of $\Delta S_{gap}(i,t) \leq 0.05$ is equal to 80%, which means that the synthetic data holds a spatial correlation property.

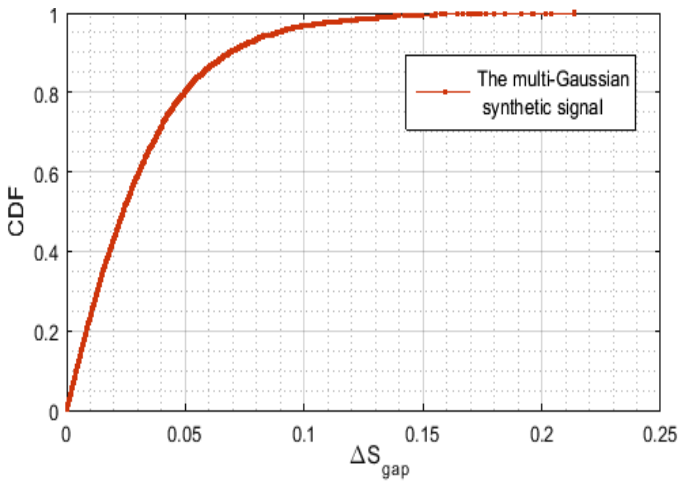


Fig. B.1. The CDF of ΔS_{gap} of a multi-Gaussian synthetic signal generated using the values of Table. 4.1.

B.2 Temporal Correlation feature

To evaluate the temporal stability of the data X, we measure the normalized gap between each two consecutive gathered data samples, $X(i,t-1)$ and $X(i,t)$, in a specific space location i. That is:

$$\Delta T_{gap}(i,t) = \frac{\mid X(i,t) - X(i,t-1) \mid}{\max\limits_{1 \leqslant i \leqslant N, 2 \leqslant t \leqslant T} \mid X(i,t) - X(i,t-1) \mid}. \tag{B.2}$$

Figure B.2 plots the CDF of the ΔT_{gap} values. We can note that the probability

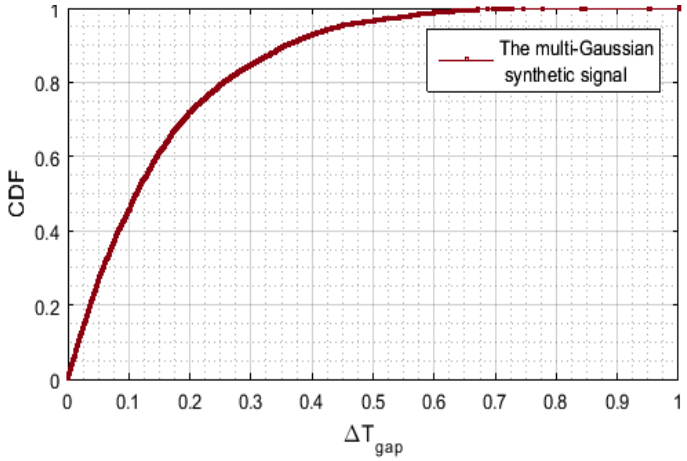


Fig. B.2. The CDF of ΔT_{gap} of a multi-Gaussian synthetic signal generated using the values of Table. 4.1.

of $\Delta T_{gap}(i,t) \leq 0.25$ is equal to 80%, which means that the synthetic data holds a temporal correlation property.

B.3 Cross Configuration

In figure B.3, we have performed a cross configuration for an empirical choice of the used tuning parameters of (4.15). As we can note from the simulation, adjusting these parameters nicely enhances the data reconstruction performance of the proposed approach. The combination ($fac_1 = 10^{-13}$, $fac_2 = 1$ and $K = 5$) seems to afford sufficiently good results compared to other tested values. Note that tuning these parameters serves just to further improve and refine the data reconstruction quality. Indeed even with the the extreme values ($fac_1 = 1$, $fac_2 = 1$ and $K = 2$), our proposed approach still achieves a very low data recovery error.

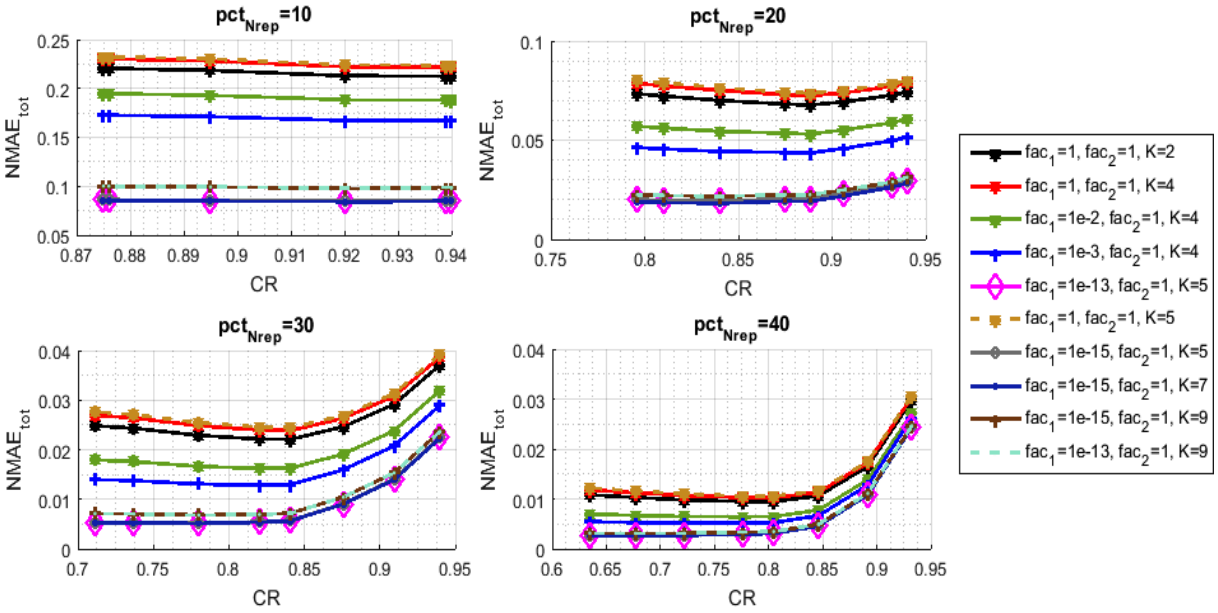


Fig. B.3. The $NMAE_{tot}$ for the proposed technique with the variation of the parameters K, fac_1 and fac_2.

Figure B.4 shows that our proposed approach ,executed without the regularization of (K, fac_1 and fac_2), still distinctly outperforms the Benchmark scheme, implemented in Figure 4.6.

In Figures B.5 and B.6, we have varied the size of the data matrix X (i.e. N and T). As we can note from these plots, the reconstruction performance, with and without parameters scaling, is independent to the performed numbers N and T.

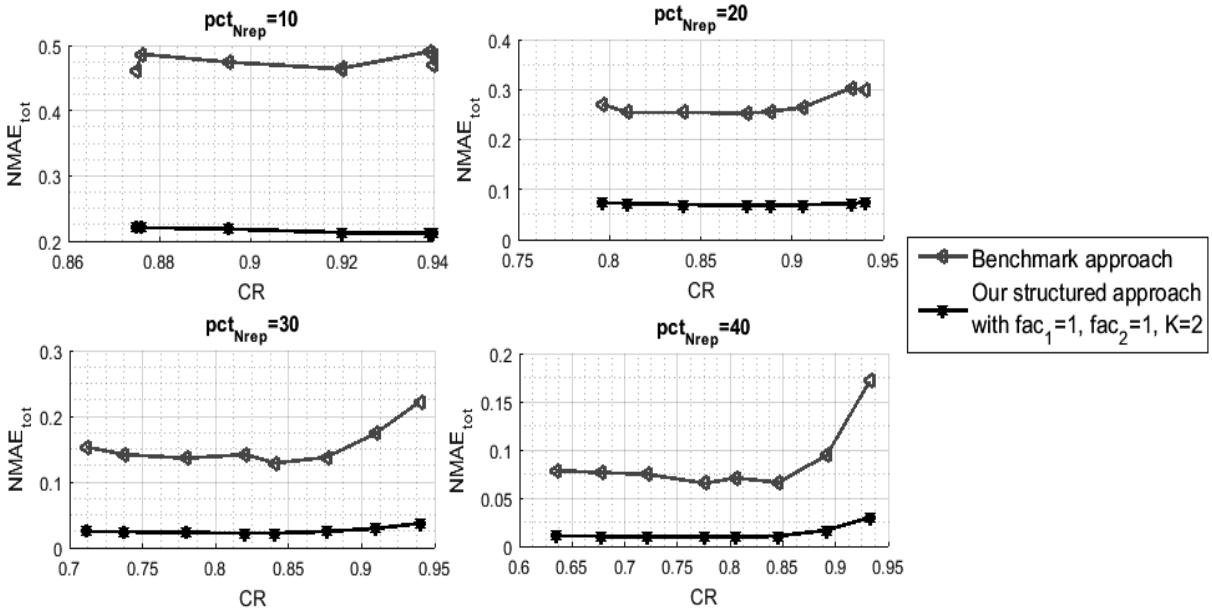


Fig. B.4. The $NMAE_{tot}$ for the Benchmark technique and for the proposed one without parameters adjustment.

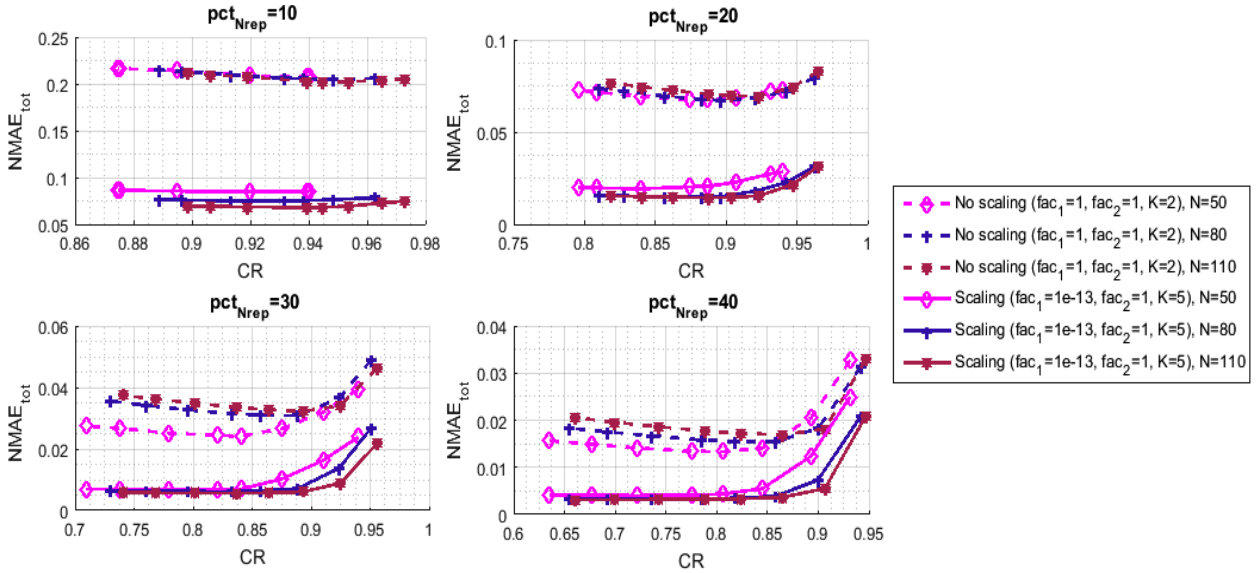


Fig. B.5. The $NMAE_{tot}$ for the proposed approach with and without parameters adjustment with respect to the number of sensor nodes N.

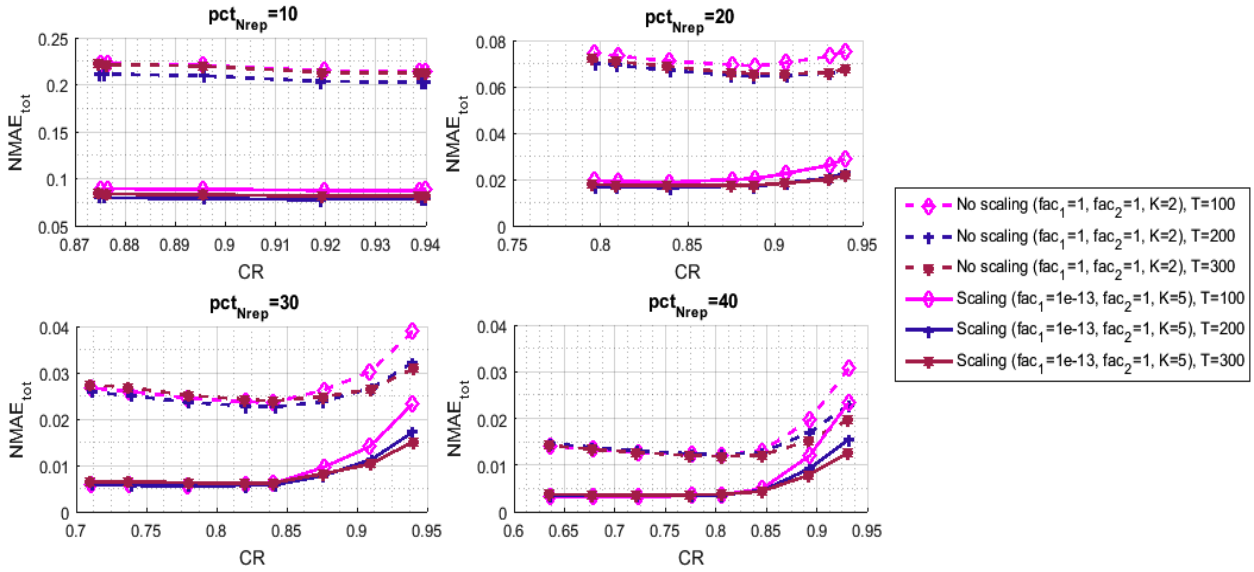


Fig. B.6. The $NMAE_{tot}$ for the proposed approach with and without parameters adjustment with respect to the number of time slots T.

Bibliography

[1] IDC. The growth in connected iot devices is expected to generate 79.4zb of data in 2025, according to a new idc forecast. 18 Jun 2019.

[2] Fortune Business Insights. Market research report. nov. 2019.

[3] Pawan Kumar Verma, Rajesh Verma, Arun Prakash, Ashish Agrawal, Kshirasagar Naik, Rajeev Tripathi, Maazen Alsabaan, Tarek Khalifa, Tamer Abdelkader, and Abdulhakim Abogharaf. Machine-to-machine (m2m) communications: A survey. *Journal of Network and Computer Applications*, 66:83–105, 2016.

[4] Godfrey Anuga Akpakwu, Bruno J Silva, Gerhard P Hancke, and Adnan M Abu-Mahfouz. A survey on 5g networks for the internet of things: Communication technologies and challenges. *IEEE Access*, 6:3619–3647, 2017.

[5] Waleed Ejaz, Alagan Anpalagan, Muhammad Ali Imran, Minho Jo, Muhammad Naeem, Saad Bin Qaisar, and Wei Wang. Internet of things (iot) in 5g wireless communications. *IEEE Access*, 4:10310–10314, 2016.

[6] Yasir Mehmood, Farhan Ahmad, Ibrar Yaqoob, Asma Adnane, Muhammad Imran, and Sghaier Guizani. Internet-of-things-based smart cities: Recent advances and challenges. *IEEE Communications Magazine*, 55(9):16–24, 2017.

[7] Mohammad Abdur Razzaque and Simon Dobson. Energy-efficient sensing in wireless sensor networks using compressed sensing. *Sensors*, 14(2):2822–2859, 2014.

[8] Malka N Halgamuge, Moshe Zukerman, Kotagiri Ramamohanarao, and Hai L Vu. An estimation of sensor energy consumption. *Progress in Electromagnetics Research*, 12:259–295, 2009.

[9] Giuseppe Anastasi, Marco Conti, Mario Di Francesco, and Andrea Passarella. Energy conservation in wireless sensor networks: A survey. *Ad hoc networks*, 7(3):537–568, 2009.

[10] S Sandeep Pradhan and Kannan Ramchandran. Distributed source coding using syndromes (discus): Design and construction. *IEEE Transactions on Information Theory*, 49(3):626–643, 2003.

[11] Tom Schoellhammer, Ben Greenstein, Eric Osterweil, Michael Wimbrow, and Deborah Estrin. Lightweight temporal compression of microclimate datasets. 2004.

[12] Sunyong Kim, Chiwoo Cho, Kyung-Joon Park, and Hyuk Lim. Increasing network lifetime using data compression in wireless sensor networks with energy harvesting. *International Journal of Distributed Sensor Networks*, 13(1):1550147716689682, 2017.

[13] Emmanuel J Candès and Michael B Wakin. An introduction to compressive sampling [a sensing/sampling paradigm that goes against the common knowledge in data acquisition]. *IEEE signal processing magazine*, 25(2):21–30, 2008.

[14] Yonina C Eldar and Gitta Kutyniok. *Compressed sensing: theory and applications*. Cambridge University Press, 2012.

[15] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[16] Leinonen Markus. *DISTRIBUTED COMPRESSED DATA GATHERING IN WIRELESS SENSOR NETWORKS*. PhD thesis, 2018.

[17] Davide Brunelli and Carlo Caione. Sparse recovery optimization in wireless sensor networks with a sub-nyquist sampling rate. *Sensors*, 15(7):16654–16673, 2015.

[18] Marco F Duarte, Michael B Wakin, Dror Baron, and Richard G Baraniuk. Universal distributed sensing via random projections. In *Proceedings of the 5th international conference on Information processing in sensor networks*, pages 177–185. ACM, 2006.

[19] Yin Zhang, Matthew Roughan, Walter Willinger, and Lili Qiu. Spatio-temporal compressive sensing and internet traffic matrices. In *ACM SIGCOMM Computer Communication Review*, volume 39, pages 267–278. ACM, 2009.

[20] Andrianiaina Ravelomanantsoa. *Approche déterministe de l'acquisition comprimée et la reconstruction des signaux issus de capteurs intelligents distribués*. PhD thesis, 2015.

[21] Mohsen Hooshmand, Michele Rossi, Davide Zordan, and Michele Zorzi. Covariogram-based compressive sensing for environmental wireless sensor networks. *IEEE Sensors Journal*, 16(6):1716–1729, 2015.

[22] Manel Kortas, Vahid Meghdadi, Ammar Bouallegue, Tahar Ezzeddine, Oussama Habachi, and Jean-Pierre Cances. Routing aware space-time compressive sensing for wireless sensor networks. In *Personal, Indoor, and Mobile Radio Communications (PIMRC), 2017 IEEE 28th Annual International Symposium on*, pages 1–6. IEEE, 2017.

[23] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

[24] Kun Xie, Xueping Ning, Xin Wang, Dongliang Xie, Jiannong Cao, Gaogang Xie, and Jigang Wen. Recover corrupted data in sensor networks: A matrix completion solution. *IEEE Transactions on Mobile Computing*, 16(5):1434–1448, 2017.

[25] Manel Kortas, Oussama Habachi, Ammar Bouallegue, Vahid Meghdadi, Tahar Ezzedine, and Jean-Pierre Cances. Energy efficient data gathering schema for wireless sensor network: A matrix completion based approach. In *Software, Telecommunications and Computer Networks (SoftCOM), 2019 International Conference*, pages 1–6. IEEE, Sept. 2019.

[26] Manel Kortas, Oussama Habachi, Ammar Bouallegue, Vahid Meghdadi, Tahar Ezzedine, and Jean Pierre Cances. The energy-aware matrix completion-based data gathering scheme for wireless sensor networks. volume 8, pages 30772–30788. IEEE, 2020.

[27] Sungwon Lee, Sundeep Pattem, and Maheswaran Sathiamoorthy. Compressed sensing and routing in multi-hop networks. Technical report, University of Southern California, 2009.

[28] Marco F Duarte and Richard G Baraniuk. Kronecker compressive sensing. *IEEE Transactions on Image Processing*, 21(2):494–504, 2012.

[29] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[30] Emmanuel Candes and Justin Romberg. l_1-magic: Recovery of sparse signals via convex programming. *URL: www. acm. caltech. edu/l1magic/downloads/l1magic. pdf*, 4:14, 2005.

[31] T Tony Cai and Lie Wang. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information Theory*, 57(7):4680–4688, 2011.

[32] H Mohimani, M Babaie-Zadeh, I Gorodnitsky, and C Jutten. Sparse recovery using smoothed l^0 (sl0): Convergence analysis, 2010. *arXiv preprint cs.IT/1001.5073*.

[33] Stephen Becker, Jérôme Bobin, and Emmanuel J Candès. Nesta: A fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, 2011.

[34] Markus Leinonen, Marian Codreanu, and Markka Juntti. Compressed acquisition and progressive reconstruction of multi-dimensional correlated data in wireless sensor networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 6449–6453. IEEE, 2014.

[35] Simon Foucart and Holger Rauhut. A mathematical introduction to compressive sensing. *Bull. Am. Math*, 54:151–165, 2017.

[36] Giorgio Quer, Riccardo Masiero, Daniele Munaretto, Michele Rossi, Joerg Widmer, and Michele Zorzi. On the interplay between routing and signal representation for compressive sensing in wireless sensor networks. In *Information Theory and Applications Workshop, 2009*, pages 206–215. IEEE, 2009.

[37] Haifeng Zheng, Feng Yang, Xiaohua Tian, Xiaoying Gan, Xinbing Wang, and Shilin Xiao. Data gathering with compressive sensing in wireless sensor networks: a random walk based approach. *IEEE Transactions on Parallel and Distributed Systems*, 26(1):35–44, 2015.

[38] Wei Wang, Minos Garofalakis, and Kannan Ramchandran. Distributed sparse random projections for refinable approximation. In *Proceedings of the 6th international conference on Information processing in sensor networks*, pages 331–339. ACM, 2007.

[39] Hossein Mamaghanian, Nadia Khaled, David Atienza, and Pierre Vandergheynst. Compressed sensing for real-time energy-efficient ecg compression on wireless body sensor nodes. *IEEE Transactions on Biomedical Engineering*, 58(9):2456–2466, 2011.

[40] Andrianiaina Ravelomanantsoa, Hassan Rabah, and Amar Rouane. Compressed sensing: A simple deterministic measurement matrix and a fast recovery algorithm. *IEEE Transactions on Instrumentation and Measurement*, 64(12):3405–3413, 2015.

[41] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[42] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[43] Zaiwen Wen, Wotao Yin, and Yin Zhang. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation*, 4(4):333–361, 2012.

[44] Matthew Roughan, Yin Zhang, Walter Willinger, and Lili Qiu. Spatio-temporal compressive sensing and internet traffic matrices. *IEEE/ACM Transactions on Networking (ToN)*, 20(3):662–676, 2012.

[45] Mark A Davenport and Justin Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.

[46] E Candes and Y Plan. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. to appear. *IEEE Trans. Info. Theo*, 2009.

[47] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *arXiv preprint arXiv:0903.1476*, 2009.

[48] Michalis Giannopoulos, Sofia Savvaki, Grigorios Tsagkatakis, and Panagiotis Tsakalides. Application of tensor and matrix completion on environmental sensing data. In *ESANN*, 2017.

[49] Yudong Chen and Yuejie Chi. Harnessing structures in big data via guaranteed low-rank matrix estimation. *arXiv preprint arXiv:1802.08397*, 2018.

[50] Chong Luo, Feng Wu, Jun Sun, and Chang Wen Chen. Compressive data gathering for large-scale wireless sensor networks. In *Proceedings of the 15th annual international conference on Mobile computing and networking*, pages 145–156. ACM, 2009.

[51] Giorgio Quer, Riccardo Masiero, Gianluigi Pillonetto, Michele Rossi, and Michele Zorzi. Sensing, compression, and recovery for wsns: Sparse signal modeling and monitoring framework. *IEEE Transactions on Wireless Communications*, 11(10):3447–3461, 2012.

[52] Zichong Chen, Juri Ranieri, Runwei Zhang, and Martin Vetterli. Dass: Distributed adaptive sparse sensing. *IEEE Transactions on Wireless Communications*, 14(5):2571–2583, 2015.

[53] Chong Luo, Feng Wu, Jun Sun, and Chang Wen Chen. Efficient measurement generation and pervasive sparsity for compressive data gathering. *IEEE Transactions on Wireless Communications*, 9(12):3728–3738, 2010.

[54] Shancang Li, Li Da Xu, and Xinheng Wang. Compressed sensing signal and data acquisition in wireless sensor networks and internet of things. *IEEE Transactions on Industrial Informatics*, 9(4):2177–2186, 2012.

[55] Jie Cheng, Qiang Ye, Hongbo Jiang, Dan Wang, and Chonggang Wang. Stcdg an efficient data gathering algorithm based on matrix completion for wireless sensor networks. *IEEE Transactions on Wireless Communications*, 12(2):850–861, 2013.

[56] Alexandros Fragkiadakis, Ioannis Askoxylakis, and Elias Tragos. Joint compressed-sensing and matrix-completion for efficient data collection in wsns. In *Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), 2013 IEEE 18th International Workshop on*, pages 84–88. IEEE, 2013.

[57] Donghao Wang, Jiangwen Wan, Zhipeng Nie, Qiang Zhang, and Zhijie Fei. Efficient data gathering methods in wireless sensor networks using gbtr matrix completion. *Sensors*, 16(9):1532, 2016.

[58] 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.

[59] Kun Xie, Lele Wang, Xin Wang, Gaogang Xie, and Jigang Wen. Low cost and high accuracy data gathering in wsns with matrix completion. *IEEE Transactions on Mobile Computing*, 17(7):1595–1608, 2018.

[60] Jiawei Tan, Wei Liu, Tian Wang, Neal N Xiong, Houbing Song, Anfeng Liu, and Zhiwen Zeng. An adaptive collection scheme-based matrix completion for data gathering in energy-harvesting wireless sensor networks. *IEEE Access*, 7:6703–6723, 2019.

[61] Jiawei Tan, Wei Liu, Mande Xie, Houbing Song, Anfeng Liu, Ming Zhao, and Guoping Zhang. A low redundancy data collection scheme to maximize lifetime using matrix completion technique. *EURASIP Journal on Wireless Communications and Networking*, 2019(1):5, 2019.

[62] Zhetao Li, YuXin Liu, Ming Ma, Anfeng Liu, Xiaozhi Zhang, and Gungming Luo. Msdg: A novel green data gathering scheme for wireless sensor networks. *Computer Networks*, 142:223–239, 2018.

[63] J Srimathi and V Valli Mayil. Fuzzy gene optimized reweight boosting classification for energy efficient data gathering in wsn.

[64] Zhetao Li, Yuxin Liu, Anfeng Liu, Shiguo Wang, and Haolin Liu. Minimizing convergecast time and energy consumption in green internet of things. *IEEE Transactions on Emerging Topics in Computing*, 2018.

[65] Yanjun Yao, Qing Cao, and Athanasios V Vasilakos. Edal: An energy-efficient, delay-aware, and lifetime-balancing data collection protocol for heterogeneous wireless sensor networks. *IEEE/ACM Transactions on Networking (TON)*, 23(3):810–823, 2015.

[66] Naeem Jan, Nadeem Javaid, Qaisar Javaid, Nabil Alrajeh, Masoom Alam, Zahoor Ali Khan, and Iftikhar Azim Niaz. A balanced energy-consuming and hole-alleviating algorithm for wireless sensor networks. *IEEE Access*, 5:6134–6150, 2017.

[67] Zahid Wadud, Nadeem Javaid, Muhammad Khan, Nabil Alrajeh, Mohamad Alabed, and Nadra Guizani. Lifetime maximization via hole alleviation in iot enabling heterogeneous wireless sensor networks. *Sensors*, 17(7):1677, 2017.

[68] K Mahendrababu and K Lakshmi Joshitha. A solution to energy hole problem in wireless sensor networks using witricity. In *International Conference on Information Communication and Embedded Systems (ICICES2014)*, pages 1–6. IEEE, 2014.

[69] G Edwin Prem Kumar, K Baskaran, R Elijah Blessing, and M Lydia. A comprehensive review on the impact of compressed sensing in wireless sensor networks. *International Journal on Smart Sensing and Intelligent Systems*, 9(2), 2016.

[70] Adriana Schulz, Eduardo Antônio Barros Da Silva, and Luiz Velho. *Compressive sensing*, volume 13. IMPA, 2009.

[71] Yong Wang, Zhuoshi Yang, Jianpei Zhang, Feng Li, Hongkai Wen, and Yiran Shen. Cs²-collector: A new approach for data collection in wireless sensor networks based on two-dimensional compressive sensing. *Sensors*, 16(8):1318, 2016.

[72] Mohammadreza Balouchestani, Kaamran Raahemifar, and Sridhar Krishnan. Robust wireless sensor networks with compressed sensing theory. In *International Conference on Networked Digital Technologies*, pages 608–619. Springer, 2012.

[73] Mohammadreza Balouchestani, Kaamran Raahemifar, and Sridhar Krishnan. Compressed sensing in wireless sensor networks: Survey. *Canadian Journal on Multimedia and Wireless Networks*, 2(1):1–4, 2011.

[74] Celalettin Karakus, Ali Cafer Gurbuz, and Bulent Tavli. Analysis of energy efficiency of compressive sensing in wireless sensor networks. *IEEE Sensors Journal*, 13(5):1999–2008, 2013.

[75] Piyush Gupta and Panganmala R Kumar. The capacity of wireless networks. *IEEE Transactions on information theory*, 46(2):388–404, 2000.

[76] Bartosz Musznicki, Mikolaj Tomczak, and Piotr Zwierzykowski. Dijkstra-based localized multicast routing in wireless sensor networks. In *Communication Systems, Networks & Digital Signal Processing (CSNDSP), 2012 8th International Symposium on*, pages 1–6. IEEE, 2012.

[77] Davide Zordan, Giorgio Quer, Michele Zorzi, and Michele Rossi. Modeling and generation of space-time correlated signals for sensor network fields. In *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*, pages 1–6. IEEE, 2011.

[78] Petter Abrahamsen. A review of gaussian random fields and correlation functions, 1997.

[79] TE Smith. Notebook on spatial data analysis. Philadelphia: University of Pennsylvania Press, 2014.

[80] Apoorva Jindal and Konstantinos Psounis. Modeling spatially correlated data in sensor networks. *ACM Transactions on Sensor Networks (TOSN)*, 2(4):466–499, 2006.

[81] Lutz Bierl. Msp430 family mixed-signal microcontroller application reports. 2000.

[82] Manel Kortas, Ammar Bouallegue, Tahar Ezzeddine, Vahid Meghdadi, Oussama Habachi, and Jean-Pierre Cances. Compressive sensing and matrix completion in wireless sensor networks. In *Internet of Things, Embedded Systems and Communications (IINTEC), 2017 International Conference on*, pages 9–14. IEEE, 2017.

[83] Chih-Chieh Hung, Wen-Chih Peng, Wang-Chien Lee, et al. Energy-aware set-covering approaches for approximate data collection in wireless sensor networks. *IEEE Transactions on Knowledge and Data Engineering*, 24(11):1993, 2012.

[84] Rong Du, Cailian Chen, Bo Yang, Ning Lu, Xinping Guan, and Xuemin Shen. Effective urban traffic monitoring by vehicular sensor networks. *IEEE Transactions on Vehicular Technology*, 64(1):273–286, 2015.

[85] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[86] Huibin Zhou, Dafang Zhang, and Kun Xie. Accurate traffic matrix completion based on multi-gaussian models. *Computer Communications*, 102:165–176, 2017.

[87] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

[88] M Grant and S Boyd. Cvx: Matlab software for disciplined convex programming, version 2.1. 2014 mar. http://www.cvxr.com/cvx,March2014.

[89] Jae-Hwan Chang and Leandros Tassiulas. Maximum lifetime routing in wireless sensor networks. *IEEE/ACM Transactions on networking*, 12(4):609–619, 2004.

[90] Divya Sharma, Sandeep Verma, and Kanika Sharma. Network topologies in wireless sensor networks: a review 1. 2013.

[91] Stts751 description. https://www.st.com/content/st_com/en/products/mems-and-sensors/temperature-sensors/stts751.html. Accessed: 2019-09-30.

[92] Siguang Chen, Zhihao Wang, and Kewei Sha. Cluster-aware kronecker supported data collection for sensory data. In *2018 27th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–6. IEEE, 2018.

[93] Zhen Gao, Linglong Dai, Wei Dai, Byonghyo Shim, and Zhaocheng Wang. Structured compressive sensing-based spatio-temporal joint channel estimation for fdd massive mimo. *IEEE Transactions on Communications*, 64(2):601–617, 2015.

[94] Bin Lv, Zhen Yang, and Youhong Feng. Temporally and spatially correlated uplink channel estimation for massive mimo systems. In *2015 International Conference on Wireless Communications & Signal Processing (WCSP)*, pages 1–5. IEEE, 2015.

[95] Mahmoud A Albreem, Markku Juntti, and Shahriar Shahabuddin. Massive mimo detection techniques: a survey. *IEEE Communications Surveys & Tutorials*, 21(4):3109–3132, 2019.

[96] Imran Khan, Joel JPC Rodrigues, Jalal Al-Muhtadi, Muhammad Irfan Khattak, Yousaf Khan, Farhan Altaf, Seyed Sajad Mirjavadi, and Bong Jun Choi. A robust channel estimation scheme for 5g massive mimo systems. *Wireless Communications and Mobile Computing*, 2019, 2019.

[97] Sinh Le Hong Nguyen and Ali Ghrayeb. Compressive sensing-based channel estimation for massive multiuser mimo systems. In *2013 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 2890–2895. IEEE, 2013.