

Review

Not peer-reviewed version

A Data-Driven Approach to Assess and Improve the Quality Of Information Service in Wireless Sensor Networks: A Survey

[Samy Benhoussa](#)*, [Gil De Sousa](#), [Jean-Pierre Chanet](#)

Posted Date: 15 January 2024

doi: 10.20944/preprints202401.1144.v1

Keywords: Internet of Thing; Wireless Sensor Networks; Quality of Service; Quality of Information; Data Reduction; Data Quality; Energy Efficiency



Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Article

A Data-Driven Approach to Assess and Improve the Quality Of Information Service in Wireless Sensor Networks: A Survey

Samy Benhoussa ^{1,*} , Gil De Sousa ² and Jean-Pierre Chanet ³

¹ Clermont Auvergne University, INRAE, TSCF Research Unit; samy.benhoussa@inrae.fr

² Clermont Auvergne University, INRAE, TSCF Research Unit; gil.de-sousa@inrae.fr

³ Clermont Auvergne University, INRAE, TSCF Research Unit; jean-pierre.chanet@inrae.fr

* Correspondence: samy.benhoussa@inrae.fr;

Abstract: One of the challenges of deploying Wireless Sensor Networks (WSN) at large scale is to operate for prolonged periods of time relying on the limited amounts of energy stored in its nodes. Generally, data communication is the main factor of draining energy. Consequently, reducing the amount of transmitted data by the nodes is of great interest. At the same time, reducing data to be transmitted should not decrease the quality of the deduced information. Solving this problematic is what we call in this article "Ensuring the Quality of Information Service (QoIS)". This document provides a state-of-the-art for the QoIS regarding WSNs. At first, we give a quick overview of the Internet of Things (IoT) paradigm and the WSN technologies in order to understand how data is gathered and used in these networks. Also, we introduce a novel concept: QoIS, in contrast with classic QoS. Then, it is expected a summary of criteria that asses the QoIS in WSNs plus some techniques to improve it. In this article, we will focus on data management techniques. These techniques can improve raw data, information and/or knowledge quality in energy-constrained WSNs. Finally, we give some conclusions and a glimpse about our future work.

Keywords: Internet of Things; Wireless Sensor Networks; Quality of Service; Data Quality; Quality of Information; Data Reduction; Energy Efficiency.

1. Introduction

Internet of Things (IoT) is a rapidly growing paradigm that is transforming the way we interact with the world around us. Many fields have been impacted positively with the emergence of IoT : industry with manufacturing industry 4.0 [1], logistics with inventory management [2], health care [3], and agriculture [4] to name few. This novel concept involves the pervasive presence of a variety of objects, such as RFID tags, sensors, actuators, and mobile phones, which are able to interact with each other and cooperate to achieve common goals. [5] [6]

Environmental science and agronomy are the particular fields we are focusing on in our researches. Agriculture 4.0 takes advantage of all the new technologies enabled by the IoT paradigm like Wireless Sensor Networks.

Other application domains of the new technologies using IoT techniques include smart homes, smart cities, and environmental monitoring. In addition to that some of the futuristic applications such as smart shopping, smart mobility, and smart governance begin to be discussed in the IoT research atmosphere [6].

Wireless sensor networks (WSNs) are a concept that has been made viable by the convergence of micro-electro/mechanical systems technology, wireless communications, and digital electronics. A wireless sensor network is a collection of small, low-power devices called sensors that are capable of sensing and transmitting data wirelessly [7]. These sensors can be deployed in various environments to monitor physical or environmental conditions such as temperature, humidity, pressure, and motion. The data collected by these sensors can be used for a wide range of applications such as environmental monitoring, military surveillance, health care, and more [1] [2] [3] [4] [7].

In environmental monitoring like many other applications, WSNs are used to collect, send, and process information in a distributed manner. These sensors are equipped with sensing and communication capabilities that allow them to collect data and transmit it to a central location for processing.

The data collected by the sensors is also processed in a distributed manner, with each sensor node performing some local processing before transmitting the data to its neighbors or the central location. This allows for efficient use of network resources and reduces the amount of data that needs to be transmitted over the network.

In addition to collecting and processing data, WSNs can also be used for control applications. For example, sensors can be used to monitor environmental conditions such as temperature or humidity and trigger actuators such as heaters or fans to maintain desired conditions.

Overall, WSNs provide a flexible and scalable platform for collecting, sending, and processing information.

1.1. Our approach and goals

Our work will focus on the Quality of Information Service (QoIS) of the WSNs. Since the WSNs primarily operate as data-centric networks, with their key emphasis placed on the valuable data generated by sensor nodes, we provide a novel definition of QoS based on the data provided by the sensor network in the data collecting phase in addition to the information transmitted to the sink in the data transmission phase. More specifically, we define the Quality of Information Service (QoIS) as the quality of the delivery of useful information by the sensor network. The quality we are discussing will take into account different criteria for evaluating the usefulness of the information provided as well as the cost of its delivery, mostly on an energetic basis. The purpose of a reasonable QoIS is to have the best information quality (or Quality of Information, QoI) at the least energetic expense. The context awareness related to IoT technologies, can be used to improve energy efficiency in WSNs by adapting sensor behavior based on contextual information [8]. For example, detecting anomalies efficiently and trigger appropriate responses [9].

Many works have explored the intersection of IoT/WSN and Data-driven optimisation techniques, such as Data Management within mHealth Environments [10], or Joint IoT/ML (Machine Learning) Platforms for Smart Societies [11].

In our paper we focus, from a data-driven point of view, on optimizing the energy consumption without altering the overall informational yield of the studied sensor network. This work will provide the most popular techniques found in the literature to guarantee this optimization, relying on Data Reduction techniques in WSNs. We intend to generalize WSNs and their different use cases. We adopt an encompassing approach where we focus only on data in WSNs no matter which type it is nor in which use case it is deployed.

1.2. The QoS and the QoIS of a WSN

There are different ways to view QoS. It could be the ability of a network to provide different levels of service to different types of traffic or applications including factors such as bandwidth, latency, packet loss, and other metrics that affect the performance and reliability of network services [12], the application reliability [13], the coverage or exposure [14] [15], which refers to covering the desired area of interest or leaving no sensing holes enabling the sensors to detect unexpected events as quickly and reliably as possible.

In our research work, we provide a novel approach of viewing the QoS; we introduce the QoIS as a concept that can measure the usefulness of information provided by a WSN while minimizing the energy cost of the data transmitting process. While other approaches are based on metrics (like bandwidth, throughput, delay etc.) to evaluate the quality of the network (information transport), we focus more on the data itself and its quality (quality of data, correctness of data, etc.) while maintaining an energy consumption threshold not to exceed. This energy threshold might be reached by the data

collecting process or the transmitting process according to the data delivery model of the WSN in question.

1.3. Data delivery models of WSNs and their QoIS

There are three main important data delivery models in WSNs :

- Event-driven model
- Query-driven model
- Continuous delivery model

These models are used to classify different classes of applications in WSNs based on their common requirements at the level of the network. These requirements are based on the user's parameterization and settings relatively to which application case is requiring the WSN data collecting. Examples of these requirements are real-time / discrete time monitoring, query-based or autonomous data sensing, etc. [13]. Authors of this latter article describes the three data delivery models of WSNs as follows:

1.3.1. Event-driven

In this model, sensor nodes detect and report events to the sink node only when they occur. The sink node is responsible for processing the event data and making decisions based on it. This model is suitable for applications that require real-time event detection and response, such as intrusion detection, fire detection, and earthquake monitoring.

1.3.2. Query-driven

In this model, the sink node sends queries to the sensor nodes to request specific data. The sensor nodes then collect and send the requested data back to the sink node. This model is suitable for applications that require periodic or on-demand data collection, such as environmental monitoring or traffic monitoring.

1.3.3. Continuous delivery

In this model, sensor nodes continuously collect and send data to the sink node without being prompted by any event or query. This model is suitable for applications that require continuous monitoring of a particular environment or system parameter, such as temperature monitoring in a greenhouse or structural monitoring of a bridge.

Each of these models has its own advantages and disadvantages. In terms of QoIS assessing and improving, requirements are not the same in every data delivery model specially when it comes to time and energy constraints.

Table 1 gives a classification of WSN data delivery models based on their perception/transmission of data. As presented in the table, we can conclude that energetic concerns are much more important in continuous delivery than in query-driven model since data in a continuous delivery model should be transmitted frequently (high frequency or even real time transmissions). In these models, guaranteeing QoIS is to reduce the weight of data to be transmitted without altering its quality of information. In event-driven model, data is perceived at a high frequency (or real time) but only transmitted if an event occurs, for these models, the accuracy of the perception of the nodes is of high concern and the energy constraint will hinder this accuracy, guaranteeing QoIS in these models is to find the perfect balance between the quality of perceived data and the energy cost of its perception. Additionally, This model needs a distributed knowledge so that nodes that perceive data can process it locally, to know if an event is occurring, without transmitting it to the sink in order to do the processing. In query-driven models, the user is the trigger of perceiving and transmitting. The user can require high quality or low quality data from the WSN, relying on their needs and the remaining energy in the WSN.

Table 1. Perception and transmission of different data-delivery models of WSNs

Data delivery model	Perception	Transmission
Event-driven	High	Low
Query-driven	Low	Low
Continuous delivery	High	High

This survey article is presented as follows : Section II will give the different criteria in assessing the Information value of data collected by sensors, Section III will give strategies and techniques to improve the QoIS based on the Information value and the energetic cost of a wireless sensor in the data collecting phase, and Section IV will give a conclusion and the possible future work based on this survey. Finally, we give acknowledgement to the funders and contributors.

2. QoIS Assessment Criteria

We conceptualize the life cycle of data (as delivered by a WSN) as follows :

- Raw data is collected by the sensor nodes respectively at the area of interest where the phenomenon is happening (example: a numeric value of temperature).
- Information which is the aggregation of raw data sensed by each node all over the network. the goal of the aggregation process is to extract meaningful information from raw data via function(s) that we name "aggregation function" (example: the mean temperature in an area, the aggregation function is the mathematical mean function).
- Knowledge retrieved from information gathered by a sensor network through out a long enough period of time to make conclusions based on these collected information and other contextual information. These conclusions are what we name "knowledge" (example : the mean temperature of a sensed area at day/night of each season. The contextual information here is the date/time when data was collected).

Accordingly to each phase of the cycle above, data quality criteria may differ.

2.1. Raw data

At this phase of the data life cycle, importance is given to the data quality metrics relative to the precision of its sensing and its transport quality. Authors of [16] give criteria called data attributes to evaluate the quality of raw data. We can bring some examples given by articles discussing Quality of Information (QoI) in WSNs:

- Accuracy: refers to the degree to which the information provided by a WSN reflects the true state of the environment being monitored [16] .
- Timeliness: refers to how quickly information is delivered from the WSN to its intended recipients [16][17]. This is particularly important for applications that require real-time or near-real-time data.
- Completeness: refers to whether all relevant data has been collected and transmitted by the WSN [16][17]. Incomplete data can lead to inaccurate or misleading results.
- Relevance: refers in the article [16] to whether the information provided by a WSN is useful and applicable for its intended purpose. Irrelevant data can waste resources and reduce overall system performance.
- Redundancy: refers to the information overlap among different sensor nodes, and can be measured via investigating the similarity of their sensing results [18]. Data provided by a sensor node that is similar to the one provided by another node is likely to be redundant.
- Consistency: refers to whether the information provided by a WSN is internally consistent and free from contradictions or errors [16]. Inconsistent data can lead to incorrect conclusions or decisions.

- Trustworthiness: refers to whether the information provided by a WSN can be trusted by its intended recipients [16]. This includes factors such as data integrity, security, and privacy.
- Volume : defines the data volume attribute of a WSN as the size of the data set which can be used to describe the working state for a given sensor node, including data collected by the node, and meta-data describing it. [19] [17].

The authors of [20] propose 3 metrics based on the previously cited attributes: currency, availability and validity. The definitions ensure that all the parameters are interpretable and obtainable. Furthermore, the paper demonstrates the feasibility of proposed metrics by applying them to real-world data sources.

2.2. Aggregated information

There are numerous criteria in the literature to assess the quality of aggregated information in a WSN (also referred to as QoI) [16][21], we can cite a few:

- Shannon entropy: is a measure of the amount of uncertainty or randomness in a system. It was introduced by Claude Shannon in 1948 as a way to quantify the amount of information contained in a message or signal [21]. The entropy of a system is defined as the average amount of information needed to describe the state of the system. In other words, it is a measure of the amount of "surprise" or "uncertainty" associated with the system. Shannon entropy has applications in a wide range of fields, including information theory, cryptography, statistical physics, and computer science.
- User requirements responding: refers to the information needs of users in a WSN. These requirements are based on a set of attributes that define the Quality of Information (QoI) that users expect from the system [16]. The article also suggests that user requirements can be thought of as measured information based on a specific set of attributes and, notes that users in a WSN may not necessarily be human, but could also be other systems or applications that rely on the data provided by the WSN.
- Error probability: is defined as a measure of the likelihood that an error will occur in the dissemination of information acquired/extracted/transmitted in a wireless sensor network (WSN).
- Path Weakness: is a game-theoretic metric used to measure the qualitative performance of different routing mechanisms in WSNs. The approach uses qualitative performance as a QoI characteristic and uses a sensor-centric concept. According to the article, path weakness is calculated by considering the probability that an attacker can successfully launch an attack on a given path in the network. The higher the path weakness, the more vulnerable the path is to attacks and therefore, lower its QoI.
- Transient Information level: is a key metric used to define the QoI assigned to a message in a WSN. According to the article, transient information level is defined as the product of information and projected physical distance of that information from destination node. In other words, transient information level takes into account both the amount of information being transmitted and how far it needs to travel to reach its destination. This approach is relevant to QoI information transport block as attributes related to information transport such as timeliness of information are used.
- Peak Signal-to-Noise Ratio (PSNR): is a metric used to measure the quality of a reconstructed signal compared to its original form. It is commonly used in image and video processing applications to evaluate the fidelity of the reconstructed signal. However, this approach only focuses on accuracy and does not consider other attributes such as timeliness for timely arrival of information for decision making.

2.3. Knowledge

This stage of the data cycle implies historical data to make adequate conclusions for the decision-making process. The criteria we will be based on are those used in Data Mining to evaluate the learning quality. We can give as examples:

- Recall: is a metric used to evaluate the performance of a classification model in machine learning. It measures the proportion of actual positive cases that were correctly identified by the model, out of all positive cases in the data set.
- False-positive Rate: is another metric used to evaluate the performance of a classification model in machine learning. It measures the proportion of negative cases that were incorrectly classified as positive by the model, out of all negative cases in the data set.
- Overall Success Rate: is somehow a combination of both Recall and False-positive rate. In other words, it is the proportion of negative cases that were correctly classified as negative by the model as well as positive cases correctly identified by the model.

As discussed before, the data quality criteria used can differ from a data-delivery model to another. Also, some criteria might be more important in some delivery models than in others, or more important in some applicative use cases than in others, etc.

3. QoIS Improvement techniques

3.1. Data Cleaning

Improving the quality of data in a WSN can be achieved by addressing the issue of unreliable data due to numerous errors in the network, which can have a significant impact on applications that rely on this data. The article [19] proposes a set of quality indicators and metrics, including data volume, completeness, accuracy, and consistency, and describes a cleaning strategy based on these indicators. The proposed method is evaluated through simulations, and the results show that it can effectively improve the quality of data in WSNs. This data cleaning strategy relying on volume, completeness, accuracy and consistency as data quality criteria consists of two main steps:

- Identifying and measuring quality indicators and metrics. This article identifies four quality indicators: data volume, completeness, accuracy, and consistency. For each indicator, the article proposes a set of metrics that can be used to measure its value. For example, the data volume indicator can be measured by counting the number of data points collected by each node in the network.
- Applying a cleaning strategy based on these indicators taking into consideration the relationships between indicators : step, the article proposes a cleaning strategy based on these indicators. This cleaning strategy consists of three main steps:
 - Identifying errors in the data using the quality metrics.
 - Repairing or removing errors using appropriate techniques such as interpolation or outlier detection.
 - Validating the cleaned data using additional quality metrics.

The article suggests that this cleaning strategy should be applied iteratively until an acceptable level of data quality is achieved. The article also notes that care should be taken to avoid over-cleaning or under-cleaning the data, as this can lead to further errors or loss of information.

Overall, this method provides a systematic approach for improving data quality in WSNs by identifying key quality indicators and applying appropriate cleaning techniques based on these indicators.

There are two main positive correlations between two data indicators among the 4 indicators previously discussed (volume, completeness, time-related and correctness) [19] :

- Completeness is positively correlated with data volume.
- Time-related indicator is positively correlated with correctness.

Taking into account these correlations, the cleaning strategy follows these steps:

- Calculate the volume indicator of data set D
- If the volume indicator is larger than a given threshold then:
 - Clean the data set by completeness indicator: identify missing or lost data points and then use appropriate techniques such as interpolation or extrapolation to fill in the missing values. The article notes that different interpolation techniques can be used depending on the characteristics of the data set, such as linear interpolation, cubic spline interpolation, or kriging.
 - Clean the data set by time-related indicator: the algorithm is based on the assumption that the time interval between two consecutive data collections is constant. The algorithm identifies and removes outliers by comparing the time interval between two consecutive data collections with a predefined threshold value.
 - Clean the data set by correctness indicator: identify and remove outliers based on statistical analysis of the data set.

After each round of cleaning, it is important to validate the cleaned data set using additional quality metrics such as accuracy, precision, and recall. This helps ensure that new errors are not introduced during the cleaning process and that important information is not lost.

The initial two steps are employed to assess whether the cleaning process is required. The volume indicator gauges the size of the collected data. If the size is exceptionally small, it could indicate potential issues with the network, such as problems with collection or transmission modes. In such cases, the system may struggle to gather sufficient data, resulting in very low reliability for the collected information.

This cleaning strategy results in having a better data quality indicators (completeness, time-related, correctness) of the sensed data set.

The authors of [22] propose a method for outlier detection using Mahalanobis distance based on kernel principal component analysis (KPCA). KPCA calculates mappings of data points, transforming the data into another feature space and effectively separating exceptional points from normal data distribution patterns. Experimental results demonstrate that KPCA is proficient in detecting irregular values quickly and effectively.

In [23], the authors present a network outlier cleaning method based on the correction of outlier values using wavelet and distance-based dynamic time warp (DTW) outlier techniques. The cleaning process is integrated into the multi-hop data forwarding process, utilizing neighbor relationships in the hop-based routing algorithm. Experiments validate the effectiveness of this method in cleaning abnormal sensing data.

[24] leverage the radial basis function for data restoration in WSNs. Additionally, [20] propose a Kd-tree-based K-nearest neighbor (KNN) data restoration algorithm using weighted variance and weighted Euclidean distance. Using these metrics, a binary search tree for k-dimensional non-missing data is built. The weight size is inversely proportional to the data loss indicator and proportional to the variance of the indicator.

Addressing time-dependent sampling jitter, [25] focus on eliminating non-uniform sampling time series by employing linear interpolation to correct data errors. The algorithm calculates a linear function by intercepting the two preceding and subsequent data points around problematic data points in the time series. This approach aims to obtain estimates close to the true values at the correct sampling time, effectively eliminating inaccuracies in data due to node sampling jitter through regular WSN data sets sampling.

3.2. Data Selection and Transmission

Proceeding in sensory data management by selection and transmission aims to maximize the QoI of the sensory data while eliminating the redundancy under the constraint of network resources [18], which leads to an overall increase in the QoIS of the WSN.

The procedure tends to maximize the quality of decision-making by improving the data selection and transmission phases in order to select and transmit data from most reliable sensor nodes, also, by taking into account the resource constraint of the network.

The reliability of a sensor node is defined in the article as the degree to which a sensor node contributes to the classification mission (decision making phase), and can be estimated through exploring the agreement between this node and the majority of others. The authors propose metrics to estimate reliability based on the clustering results of sensor nodes. By comparing the clustering results of each node with those of its neighbors, they can calculate a reliability score for each node that reflects its contribution to the classification task. The higher the score is, the more reliable the data provided by that node is.

In other words, in a general case, the reliability of a sensor node is its contribution to the decision made by the network's decision making support system. In a scheme where sensor nodes send clustered data to the sink, it is the level to which the clustering result of the node agrees with those of the majority of other nodes.

Therefore, in order to improve QoI and reduce redundancy in data transmission, the paper proposes a data selection and transmission scheme that selects a subset of sensor nodes based on their reliability and redundancy scores and transmits their data to the sink for decision aggregation. The authors also propose metrics to estimate reliability and redundancy based on clustering results and use them to formulate an optimization problem that aims at maximizing QoI while eliminating redundancies under resource constraints.

In order to pursue the resolving of this optimisation problem, and perform the best possible selection of sensor nodes to transmit their data, the article defines the two metrics: reliability to maximize and redundancy to minimize.

The reliability metric is formulated based on the clustering results of sensor nodes. The authors propose a novel online algorithm to estimate the QoI of individual sensor nodes through exploring their clustering results reported to the sink of each mission. The reliability can be established by calculating the entropy and thus the degree of impurity of a sensor node data.

In contrast to decision aggregation, which works offline on all the collected data, sensor selection calls for a online mechanism which can adaptively determine the set of sensor nodes to transmit based on their recent data's quality of information. The criteria to assess data's quality of information in this method is reliability, which is incrementally estimated using Incremental Reliability Estimation (IRE) Algorithm.

The algorithm operates with a sliding time window that encompasses a specific number of events. QoI estimates are automatically updated as events within the window are refreshed. Additionally, each source node updates its clustering result through either re-clustering or incremental clustering. Subsequently, the clustering results of newly arrived data are reported to the sink node for each collecting mission (collecting the average temperature is a collecting mission for example) where the IRE algorithm is invoked.

The redundancy metric is formulated based on the clustering results of sensor nodes. The authors propose a metric to measure the redundancy between two sensor nodes through investigating the similarity of their clustering results.

The redundancy score between two nodes is calculated as the ratio of the number of common clusters to the total number of clusters reported by both nodes. The higher the redundancy score is, the more similar their clustering results are, and thus, the more redundant their data is.

The objective of the selection phase is to first estimate the reliability of each sensor node, then, according to available resources (the article highlights the bandwidth constraints) and data redundancy

constraint, select the most reliable sensor nodes to transmit their data to the sink. Then, the transmission phase routes and schedules the data transmitted packets to respect the bandwidth constraints.

The data selection and transmission procedure results in having a better QoI in terms of reliability and less redundancy while respecting resource constraints.

Data selection scheme can also follow a Bayesian approach [26]. The approach in this article involves selecting observations that minimize uncertainty about the variables of interest while also minimizing the required number of observations. The authors use Gaussian processes to model the data and detect change points and faults in the system. The method also takes into account the cost associated with taking observations, such as battery energy or computational cost. This is presented in the literature as the *ActiveDataSelection*. Other works discuss this topic such as [27], [28].

3.3. Data Approximation

Data approximation in WSN collect data from environments at a discrete point of time, so it gives an approximation of the continuous state. Approximate answers are good enough for most of the applications specially in the domain of environment monitoring where WSN has become an excellent solution. [29]

The article [30] discusses the challenges of collecting periodic data in WSNs, which are often battery-powered and deployed in hard-to-reach areas for prolonged periods. The authors propose a solution architecture called 2PDA (2-Phase Data Approximation). An initial version of this architecture was published in [31]. The 2PDA architecture can help extend the network's lifetime by reducing the amount of data transmitted while maintaining a high level of accuracy. In other words, the solution tends to maintain the accuracy criteria while reducing the energy consumption by reducing the amount of transmitted data.

The 2PDA architecture is based on temporal correlation between sensor readings, which allows for accurate approximation of missing data points. The authors evaluate the performance of 2PDA using simulation experiments and show that it can significantly reduce energy consumption while maintaining a high level of accuracy. The technique is based on linear regression method and very light-weight in implementation. Also, it considers application precision at each step of data approximation. If the application precision is satisfied it models the sensor data and transmits only the parameters, otherwise, data is performed without any approximation.

The work as described in the article is based on two algorithms: CORRELATED and LSM. The first one is used to determine whether there is a correlation between the sensed variable and time. If it is the case then LSM algorithm computes parameters (slope and y-intercept) to approximate the data to send and transmit the two parameters instead of the whole sensed data packet. If the first algorithms outputs that there is no correlation, raw data is sent instead with no approximation.

3.4. Data Compression

Data Compression techniques aim to reduce the amount of data transmitted over wireless channels in WSNs [32][33]. These techniques can help reduce power consumption, which is a critical problem affecting the lifetime of WSNs. By reducing the amount of data that needs to be transmitted between nodes, data compression can also help to reduce the required inter-node communication, which is the main power consumer in WSNs. This can lead to longer network lifetimes and more efficient use of limited resources such as battery power. Additionally, data compression can help to reduce network congestion and improve overall network performance by reducing the amount of data that needs to be transmitted and processed by individual nodes.

The data compression techniques as well as the aggregation techniques are used to minimize the amount of data to be transmitted over a WSN. Since radio communications between the nodes is the biggest energy consumer, limiting the cost of these communications can improve significantly the WSN lifetime in terms of energy. Data compression, by reducing the size of transmitted data over the

network, reduces the cost of communication between each node which helps improving the lifetime of the whole network.

The article provides a comprehensive review of existing data compression approaches in WSNs. Numerous data compression algorithms can be found in the literature. Many of them can be used in the case of WSNs. We can cite some of the most commonly used algorithms:

- Discrete Cosine Transform (DCT): is a mathematical technique that converts a signal or image from its spatial domain into its frequency domain. It is similar to the Fourier Transform, but it uses only real numbers and cosine functions instead of complex numbers and sine/cosine functions. The DCT is widely used in image and video compression algorithms, such as JPEG (Joint Photographic Experts Group) and MPEG (Moving Picture Experts Group), because it can efficiently represent signals with smooth variations in their frequency content [33]. The DCT coefficients can be quantized and encoded using lossy or lossless compression techniques to reduce the amount of data needed to represent an image or a video stream without significant loss in visual quality.
- Discrete wavelet transform (DWT): is a mathematical technique that decomposes a signal or an image into a set of wavelets, which are small waves that can capture localized features or details in the signal. Unlike the Fourier Transform, which uses only sine and cosine functions to represent signals in their frequency domain, wavelets can be designed to have different shapes and scales that can better capture different types of features in a signal [33]. The DWT is widely used in image and video compression algorithms, such as JPEG2000, because it can efficiently represent signals with both smooth variations and sharp edges or details in their frequency content.
- Differential Encoding: is a simple data compression technique that is commonly used in WSNs. The basic idea behind differential encoding is to encode the difference between consecutive data samples instead of encoding the actual data values themselves [34]. This can be useful in cases where the data being transmitted has a high degree of correlation or similarity between consecutive samples. In differential encoding, the first data sample is transmitted as-is, while subsequent samples are encoded as the difference between the current sample and the previous one. This difference value can be represented using fewer bits than the original sample value, resulting in a reduction in the amount of data that needs to be transmitted. At the receiver end, the original data can be reconstructed by adding up all of the difference values starting from the first sample. Differential encoding can be particularly effective for applications where small changes in sensor readings are more important than absolute values, such as temperature or humidity monitoring. However, it may not be suitable for applications where large changes in sensor readings need to be accurately captured and transmitted.
- Run-Length Encoding (RLE): is a data compression technique used in WSNs. The basic idea behind RLE is to represent long sequences of repeated data values with a single symbol or code. This can be useful in cases where the data being transmitted has long runs of identical or similar values [35]. In RLE, the data stream is scanned for runs of consecutive identical values. Each run is then replaced with a code that represents the length of the run and the value being repeated. For example, if a sequence of 10 zeros is encountered, it can be replaced with a code that represents "repeat 0 ten times". This can result in significant reductions in the amount of data that needs to be transmitted, especially for applications where long runs of identical or similar values are common. At the receiver end, the original data can be reconstructed by decoding each run code and repeating the corresponding value for the specified number of times. RLE is a simple and efficient compression technique that can be implemented using minimal computational resources, making it well-suited for use in resource-constrained WSNs. However, it may not be as effective for compressing more complex or diverse types of data.
- Huffman Coding: is a popular data compression technique that is commonly used in WSNs. The basic idea behind Huffman coding is to assign variable-length codes to different symbols based

on their frequency of occurrence in the data stream [36] [37]. Symbols that occur more frequently are assigned shorter codes, while symbols that occur less frequently are assigned longer codes. In Huffman coding, a binary tree is constructed based on the frequency of occurrence of each symbol in the data stream. The most frequent symbols are placed near the root of the tree, while less frequent symbols are placed further away from the root. Each symbol is then assigned a unique binary code based on its position in the tree. The resulting code is a prefix code, which means that no code word is a prefix of any other code word. At the receiver end, the original data can be reconstructed by decoding each code word using the same binary tree used for encoding. Huffman coding can be very effective for compressing data streams with non-uniform symbol frequencies, such as text or image data. However, constructing an optimal Huffman tree requires knowledge of the frequency distribution of symbols in advance, which may not always be available in real-world applications. The authors of [37] show that given general knowledge of the parameters that must be monitored, conventional Huffman coding can be used to compress the data collected by the sensor nodes. When the data consists of integer measurements, the Huffman dictionary computed using statistics inferred from public datasets often approaches the entropy of the data. This allows for efficient compression of data with minimal loss of information, making it a suitable method for use in WSNs.

- Lempel-Ziv-Welch (LZW) Compression: considered as a lossless data compression algorithm, it is commonly used in WSNs. The basic idea behind LZW compression is to replace frequently occurring sequences of symbols with shorter codes, thereby reducing the overall size of the data stream [38]. In LZW compression, a dictionary of symbol sequences and their corresponding codes is built up as the data stream is processed. Initially, the dictionary contains all possible single symbols in the data stream. As the data stream is processed, frequently occurring symbol sequences are added to the dictionary and assigned shorter codes. At the encoder end, each symbol sequence in the data stream is replaced with its corresponding code from the dictionary. The resulting code sequence can be transmitted using fewer bits than would be required to represent each symbol individually. At the receiver end, the original data can be reconstructed by decoding each code and using the same dictionary used for encoding. LZW compression can be very effective for compressing text or image data streams with repetitive patterns or long runs of identical symbols. However, it requires more computational resources than some other compression techniques and may not always be suitable for use in resource-constrained WSNs.
- Arithmetic Coding: the basic idea behind arithmetic coding is to represent a sequence of symbols as a single fractional value between 0 and 1, which can then be encoded using fewer bits than the original sequence [39]. In arithmetic coding, each symbol in the data stream is assigned a probability based on its frequency of occurrence. A cumulative probability distribution is then constructed by adding up the probabilities of all symbols up to and including the current symbol. The entire data stream can be represented as a single fractional value within the range defined by the cumulative probability distribution. At the encoder end, the fractional value representing the entire data stream is then converted into a binary code using fewer bits than would be required to represent each symbol individually. At the receiver end, the original data can be reconstructed by decoding each binary code and using the same probability distribution used for encoding. Arithmetic coding can be very effective for compressing data streams with non-uniform symbol frequencies, such as text or image data. However, it requires more computational resources than some other compression techniques and may not always be suitable for use in resource-constrained WSNs.
- Burrows-Wheeler Transform (BWT): is a lossless data compression technique that is commonly used in WSNs. The basic idea behind BWT is to rearrange the symbols in a data stream to create a new sequence that has more repeated patterns, which can then be compressed more effectively [40]. In BWT, the original data stream is first transformed into a matrix of all possible cyclic permutations of the symbols in the stream. The rows of this matrix are then sorted lexicographically to create a new matrix. The last column of this new matrix is then extracted

and used as the transformed data stream. At the encoder end, this transformed data stream is then compressed using techniques such as run-length encoding or Huffman coding. At the receiver end, the original data can be reconstructed by reversing the BWT process. BWT can be very effective for compressing text or image data streams with repetitive patterns or long runs of identical symbols. However, it requires more computational resources than some other compression techniques and may not always be suitable for use in resource-constrained WSNs.

- **Prediction-Based Compression:** is a data compression technique that is commonly used in WSNs. The basic idea behind prediction-based compression is to use the correlation between adjacent data samples to predict the value of the next sample, and then encode the difference between the predicted value and the actual value [41]. In prediction-based compression, a model is first created to predict the value of each sample based on its previous samples. This model can be as simple as a linear predictor that uses a weighted sum of the previous samples, or it can be more complex, such as an auto-regressive model that uses a linear combination of past samples and past prediction errors. At the encoder end, each sample in the data stream is predicted using this model, and then the difference between the predicted value and the actual value is encoded using techniques such as Huffman coding or arithmetic coding. At the receiver end, these differences are decoded and added back to the predicted values to reconstruct the original data stream. Prediction-based compression can be very effective for compressing data streams with high temporal correlation, such as audio or video data. However, it requires more computational resources than some other compression techniques and may not always be suitable for use in resource-constrained WSNs.

Lossless data compression techniques are preferred such as Lempel-Ziv-Welch, Burrows-Wheeler Transform or Prediction-Based Compression. However, the authors noted that while these techniques can be effective for compressing data streams with repetitive patterns or high temporal correlation, they may not always be suitable for use in resource-constrained WSNs due to their computational requirements.

3.5. Data Prediction

Prediction methods aim to reduce the amount of transmitted data in WSNs [42]. We can classify these methods under two categories: Single Prediction Scheme (SPS) and Dual Prediction Scheme (DPS) according to each prediction scheme category, the article makes distinction between schemes that have the model generated at the Cluster Head / Sink and the ones that have their model generated in the sensor nodes.

- **Single Prediction Scheme:**

In this category, prediction is made in a single point of the network (sensor or cluster head/ sink). For example, the Cluster Head has the capability to anticipate the data obtained from sensor nodes and independently determine the optimal timing for acquiring additional measurements, taking into account the reliability of the predictions. Alternatively, sensor nodes can forecast alterations in their environment, enabling them to circumvent unnecessary measurements (and consequently unnecessary transmissions).

- **Model generation in the Cluster Heads / Sink:**

Given that Cluster Heads (CHs) / Sinks possess greater computational power and energy resources, they can formulate sophisticated prediction models and make crucial decisions regarding the WSN's operation without compromising the QoI delivered by the measurements. In this case, the sensor nodes are tasked solely with their fundamental functions, which involve measuring environmental parameters and transmitting the raw data gathered by their sensors.

In environmental monitoring, the data sensed by nodes typically exhibit spatio-temporal correlations, facilitating the creation of probabilistic models. This allows for the approximation of data to well-known distributions and the assignment of confidence levels to predictions. As a result, the number of transmissions is minimized, as CHs predict measurements and locally verify whether the necessary QoI constraints are met or not. Examples of this model generation scheme include adaptive sampling, topology control, and clustering.

- Model generation in the Sensor Nodes:

The prediction is made at the level of the sensor nodes. Considering the potential constraints on the computing power of sensor nodes, decisions regarding predictions can be supported by the data from their neighboring nodes, making the process distributed. For example, rather than transmitting every measurement to the Cluster Head (CH), a sensor node may autonomously decide not to transmit if it observes that the measurements from its neighbors are adequate for accurately monitoring its region.

- Dual Prediction Scheme:

In this category, predictions are made in CHs and Sensor Nodes in a simultaneous way. The underlying concept of such mechanisms is that sensor nodes possess the capability to generate the same "a priori" knowledge as CHs. However, sensor nodes can independently verify the accuracy of predictions locally, thereby avoiding unnecessary transmissions.

- Model generation in Cluster Heads/Sink:

This scheme capitalizes on the asymmetric distribution of computational power in Wireless Sensor Networks (WSNs): Cluster Heads (CHs)/Sinks typically have more abundant resources, including cheaper energy sources, as well as more memory and processing power, compared to ordinary sensor nodes primarily utilized for measuring and reporting environmental data. Sensor nodes transmit their current measurements to the CHs, enabling them to locally, based on the received values, generate new prediction models for each sensor node, update and transmit new model parameters and error acceptance levels to their sensor nodes. Decisions can be made at the level of the CH, for example the Dual Kalman Filter can be employed, leveraging spatial correlations among measurements from various sensor nodes. Subsequently, the CHs assess their capacity to compute multiple prediction models for each sensor node and opt for the most suitable one based on the received measurements. Alternatively, decisions are made locally within the sensor nodes, using methods such as GP (Gaussian Process) Regression. In this context, each sensor must forecast the information it is about to sample and adapt its sampling schedule in accordance with energy constraints. The goal is to maximize the information collected during a specific time interval.

- Independent model generation:

This scheme relies on an "initialization phase," which is a designated period during which sensor nodes report all the data they have generated to CHs. The purpose of the initialization phase is to ensure that CHs possess comprehensive information about the environment before any prediction model is generated. Following the initialization, CHs can generate prediction models similar to those in the sensor nodes without requiring additional transmissions. Subsequently, both entities commence predicting values, with the advantage that sensor nodes can locally verify the accuracy of predictions. If a prediction is deemed inaccurate, the sensor nodes may transmit the actual measurement as needed. Consequently, sensor nodes have the option to regularly report data to CHs in cases of prediction inaccuracies or refrain from reporting any sensor reading if the predictions are deemed sufficiently accurate.

Decisions can be made in the CH/Sink, CHs have the capability to adjust the operation of sensor nodes based on the potential savings that predictions may bring. CHs assess the

cost-effectiveness of making predictions in sensor nodes by calculating the relationship between prediction accuracy, the correlation between measurements, and the error tolerance acceptable to the user. This evaluation guides CHs in determining whether it is advantageous to implement predictions in sensor nodes or not. Otherwise, decisions can be made in the sensor nodes. Sensor nodes may possess the capability to make additional decisions contingent on the accuracy of predictions. They can autonomously choose to aggregate data received from neighboring nodes by excluding measurements that fall within their confidence interval, rather than forwarding them to the CHs.

- Model generation in sensor nodes:

Sensor nodes can generate prediction models by distributed knowledge in the WSN. It can be disseminated from one node to its neighbors, providing additional information about the surroundings enhances each node's capacity to compute more accurate prediction models before transmitting their parameters to the CH/Sink.

There are different categories of predictions models adopted in the case of data reduction in WSNs [42]. The main largely utilized models are: Time Series methods, Regression methods and Machine Learning techniques.

- Time Series methods:

A time series is a series of data points, typically comprised of observations recorded over a specific time interval and arranged in chronological order. A time series prediction model takes as input a time series to generate predictions. These predictions are expressed as a function of past observations and their corresponding time. Typically, this function is defined by parameters calculated based on previous observations. These parameters should be updated over time since the environment may evolve and change. There are many time series methods. The most popular are: the naive approaches, the Auto-regressive (AR), the Moving Average (MA), the Exponential Smoothing (ES) and the Auto-regressive Integrated Moving Average (ARIMA) [43].

- Regression methods:

Regression methods take a different approach compared to time series methods. Rather than relying solely on past values for predictions, they also predict measurements based on different types of measurements. For example, when a value is observed by one sensor node, a regressive model can be employed to predict the value that would be observed by another sensor node. The main regression methods used in WSN data reduction paradigm are: the Linear Regression, the Kernel Regression, the Gaussian process regression and the Principal Component Analysis (PCA). [44]

- Machine Learning techniques:

Machine Learning techniques have found applications in various solutions for WSNs across different levels, including routing, medium access control, and event detection. Among these solutions, Artificial Neural Networks stand out as the only ones specifically applied to reduce the number of transmissions in a WSN. [45].

Different Neural Networks can be employed to reduce transmissions in WSNs and consequently reduce energy consumption [46]. Some of these Neural Networks are Self Organizing Maps (SOM), Back Propagation (BP), Radial Basis Function (RBF)... etc.

Other data prediction methods exist in the literature such as the multi-node multi-feature (MNMF) [47], based on bidirectional long short-term memory (LSTM) network for WSNs. The proposed method considers the spatial-temporal correlation among sensory data, and aims to improve data quality and reduce unnecessary data transmission. The authors use the quartile method and wavelet threshold de-noising method to improve data quality, and then use the bidirectional LSTM neural network to learn prediction features. Experimental results are provided in [47] to demonstrate the effectiveness of the proposed MNMF method in improving prediction accuracy compared to other existing methods.

3.6. Smart Sampling

Sampling methods aim to reduce correlations in communicated data. Sampling can be based on a prediction model that relies on a neural network which performs an adaptive, data-driven, and non-uniform sampling [48]. The neural algorithm is designed to forecast sensor measurements and their uncertainties, enabling the system to minimize communications and transmitted data. Specifically, a multilayer perceptron (MLP) network is employed for this purpose.

We summarize the improvement techniques cited before, and the advantages and disadvantages of each one in the Table 2.

Table 2. QoS Improvement techniques and their advantages/disadvantages.

Method	Improved criteria	Advantages	Disadvantages	References
Data Cleaning	Depends on the cleaning scheme: completeness, correctness, volume, availability, effectiveness, accuracy, timeliness, etc.	<ul style="list-style-type: none"> Increases network performance by prioritizing high-quality data, based on the chosen criterion. Can be customized to specific data sets based on their characteristics. 	<ul style="list-style-type: none"> May require significant computational resources Effectiveness may depend on the data characteristics. May not be suitable for all types of data or networks. May require significant expertise. 	[19] [22] [23] [24] [25]
Reliability and Redundancy Selection and Transmission		<ul style="list-style-type: none"> Maximizing the reliability of sensory data. Eliminating redundancies. Maximally utilizing the network resources. 	<ul style="list-style-type: none"> Require significant computational resources to implement. May not be suitable for all sensor network applications. 	[18] [26] [27] [28]
Approximation	Accuracy, Reliability and Energy efficiency	Extends the lifetime of sensor nodes by reducing the amount of data transmitted while maintaining a high level of accuracy.	<ul style="list-style-type: none"> Lacks practical implementation. Performance under different network topologies. The impact of data loss on accuracy. 	[29] [30] [31]
Compression	Entropy, Probability of error and Energy efficiency.	<ul style="list-style-type: none"> Reduced data transmission time and energy consumption. Reduced bandwidth requirements. Increased data security and privacy. 	<ul style="list-style-type: none"> Increased computational complexity and processing time. Reduced data fidelity and quality. Limited compatibility with certain types of data. Increased implementation costs. 	[32] [33] [34] [35] [36] [37] [39] [38] [40] [41]
Prediction	Accuracy, Timeliness and Energy efficiency	<ul style="list-style-type: none"> Reduced energy consumption. Improved accuracy. Reduced network traffic. 	<ul style="list-style-type: none"> Increased complexity. Reduced precision. Limited applicability. 	[42] [43] [45] [46] [47]
Smart Sampling	Energy efficiency	<ul style="list-style-type: none"> Reduced power consumption. Reduced network congestion. Cost-effective. 	<ul style="list-style-type: none"> Reduced accuracy Limited applicability. 	[48]

4. Conclusions and Future work

In conclusion, this article has presented a novel view of Quality of Service (QoS) in Wireless Sensor Networks (WSNs) : Quality of Information Service (QoIS). Also, it surveyed data-driven approaches to assess and improve the QoIS in WSNs. We brought criteria found in the literature to estimate the QoIS, and have discussed the main challenges in deploying WSNs at a large scale. We presented some of the state-of-art techniques used to improve the QoIS such as data selection and transmission schemes, compression and prediction, techniques that aim to maximize the Quality of Information (QoI) while eliminating useless energy consumption. Furthermore, we came to the general conclusion that every

data-driven technique can lie either under the prediction scheme type or the compression type, since approximation and cleaning can be viewed as a compression technique while smart sampling and selection/transmission can be viewed as a prediction technique. We can conclude that these two types of techniques (compression and prediction) are the main techniques used to reduce the energetic cost of data acquisition and dissemination in a WSN. Data prediction aims to reduce energy costs of the perception and the transmission phases while compression techniques reduce energy costs of the transmission phase. This makes prediction more suitable for event-driven delivery models and compression more suitable for continuous delivery models of WSNs.

Exploring QoS improvement techniques and their pros and cons led us to emphasise the importance of the prediction techniques since it can decrease the amount of both perceived and transmitted data and consequently lead to a more optimal energy use. Prediction techniques based on Distributed Artificial Intelligence, by making the WSN act as a Multi-Agent System powered by a pre-trained neural network, can make the best improvement on energy consumption in a WSN especially in the application cases of environment monitoring and smart agriculture since the sensors generally only rely on limited power supply. Our future work will focus on implementing a Distributed Artificial Intelligence (DAI) in the edge of a WSN to efficiently predict data/information and avoid energy draining for a long-lasting monitoring device in an Agriculture 4.0 use case.

Author Contributions: Writing—original draft preparation, S.B.; writing—review and editing, S.B., G.DS. and J.P.C.; investigation, S.B.; supervision, G.DS. and J.P.C.; funding acquisition, G.DS. and J.P.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Agence Nationale de la Recherche (ANR) of the French Government through the program "Investissements d'Avenir" (16-IDEX-0001 CAP 20-25), Clermont Auvergne Metropole and INRAE MathNum Scientific Department.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors acknowledge the support received for the Agence Nationale de la Recherche (ANR) of the French Government through the program "Investissements d'Avenir" (16-IDEX-0001 CAP 20-25), Clermont Auvergne Metropole and INRAE MathNum Scientific Department. We also thank INRAE for providing access to research facilities and resources, support and encouragement during the writing process. We are also grateful to the anonymous reviewers who provided insightful comments and feedback on earlier versions of this manuscript. Finally, we would like to acknowledge the countless individuals whose work has contributed to the advancement of knowledge in our field. Without their contributions, this research would not have been possible.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kalsoom, T.; Ahmed, S.; Rafi-ul Shan, P.M.; Azmat, M.; Akhtar, P.; Pervez, Z.; Imran, M.A.; Ur-Rehman, M. Impact of IoT on Manufacturing Industry 4.0: A New Triangular Systematic Review. *Sustainability* **2020**, *12*, 8465.
2. Mashayekhy, Y.; Babaei, A.; Yuan, X.M.; Xue, A. Impact of Internet of Things (IoT) on Inventory Management: A Literature Survey. *Logistics* **2020**, *6*, 33.
3. Dash, S.P. The Impact of IoT in Healthcare: Global Technological Change & The Roadmap to a Networked Architecture in India. *J. Indian Inst. Sci.* **2020**, *100*, 773–785.
4. Khanna, A.; Kaur, S. Evolution of Internet of Things (IoT) and its significant impact in the field of Precision Agriculture. *Computers and Electronics in Agriculture* **2019**, *157*, 218–231.
5. Weiser, M. Some Computer Science Issues in Ubiquitous Computing. *Communications of the ACM* **1993**, *36*, 75–84. <https://doi.org/http://doi.org/10.1145/159544.159617>.
6. Atzori, L.; Iera, A.; Morabito, G. The internet of things: A survey. *Computer networks* **2010**, *54*, 2787–2805.
7. Akyildiz, I.F.; Su, W.; Sankarasubramaniam, Y.; Cayirci, E. Wireless sensor networks: a survey. *Computer networks* **2002**, *38*, 393–422.

8. Dey, A.K.; Abowd, G.D. Towards a Better Understanding of Context and Context-Awareness. In Proceedings of the in the Workshop on The What, Who, Where, When, and How of Context-Awareness, as part of the 2000 Conference on Human Factors in Computing Systems (CHI 2000), The Hague, The Netherlands, April 3, 2000, 2000.
9. Perera, C.; Zaslavsky, A.; Christen, P.; Georgakopoulos, D. Context Aware Computing for The Internet of Things: A Survey. *IEEE Communications Surveys & Tutorials* **2014**, *16*, 414–454.
10. O'Donoghue, J.; Herbert, J. Data Management within MHealth Environments: Patient Sensors, Mobile Devices, and Databases. *J. Data and Information Quality* **2012**, *4*. <https://doi.org/10.1145/2378016.2378021>.
11. Attar, H. Joint IoT/ML Platforms for Smart Societies and Environments: A Review on Multimodal Information-Based Learning for Safety and Security. *J. Data and Information Quality* **2023**, *15*. <https://doi.org/10.1145/3603713>.
12. Badis, H.; Munaretto, A.; Al Agha, K.; Pujolle, G. QoS for Ad hoc Networking Based on Multiple Metrics: Bandwidth and Delay. In Proceedings of the Proceedings of IEEE MWCN2002. IEEE, 2003, pp. 1–6.
13. Chen, D.; Varshney, P.K. QoS support in wireless sensor networks: a survey. *Wireless Communications and Mobile Computing* **2004**, *4*, 907–932.
14. Meguerdichian, S.; Koushanfar, F.; Potkonjak, M.; Srivastava, M.B. Coverage problems in wireless ad-hoc sensor networks. *ACM Sigmobile Mobile Computing and Communications Review* **2001**, *5*, 97–101.
15. Meguerdichian, S.; Koushanfar, F.; Qu, G.; Potkonjak, M. Exposure in wireless ad-hoc sensor networks. In Proceedings of the Proceedings of the 8th annual international conference on Mobile computing and networking, 2001, pp. 139–150.
16. Sachidananda, V.; Khelil, A.; Suri, N. Quality of Information in Wireless Sensor Networks: A Survey. *Dependable, Embedded Systems and Software Group* **2010**, *10*.
17. Lehner, W.; Klein, A. Representing Data Quality in Sensor Data Streaming Environments. *J. Data Inf. Qual.* **2009**, *1*, 1–28. <https://doi.org/10.1145/1577840.1577845>.
18. Su, L.; Hu, S.; Li, S.; Liang, F.; Gao, J.; Abdelzaher, T.F.; Han, J. Quality of information based data selection and transmission in wireless sensor networks. In Proceedings of the Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems, 2014, pp. 1–14.
19. Cheng, H.; Feng, D.; Shi, X.; Chen, C. Data quality analysis and cleaning strategy for wireless sensor networks. *Research Open Access* **2015**, *1*, 1–10.
20. Li, F.; Nastic, S.; Dustdar, S. Data Quality Observation in Pervasive Environments. *IEEE Transactions on Mobile Computing* **2012**, *XX*, 602–602.
21. Lesne, A. Shannon entropy: a rigorous notion at the crossroads between probability, information theory, dynamical systems and statistical physics. *Mathematical Structures in Computer Science* **2014**, *24*, e240311. <https://doi.org/10.1017/S0960129512000783>.
22. Ghorbel, O.; Ayedi, W.; Snoussi, H.; Abid, M. Fast and Efficient Outlier Detection Method in Wireless Sensor Networks. *IEEE Sensors Journal* **2015**, *15*, 3403–3411. <https://doi.org/10.1109/JSEN.2015.2388498>.
23. Zhuang, Y.; Chen, L. In-network outlier cleaning for data collection in sensor networks. In Proceedings of the Proceedings of the 2006 international conference on Wireless communications and mobile computing. ACM, 2006, pp. 1057–1062.
24. Hamrani, A.; Belaidi, I.; Monteiro, E.; Lorong, P. On the Factors Affecting the Accuracy and Robustness of Smoothed-Radial Point Interpolation Method. *Advances in Applied Mathematics and Mechanics* **2017**, *9*, 43–72.
25. Rahm, E.; Do, H.H. Data cleaning: problems and current approaches. *Bulletin of the Technical Committee on Data Engineering* **2000**, *23*, 3–13.
26. Osborne, M.A.; Garnett, R.; Roberts, S.J. Active data selection for sensor networks with faults and changepoints. *IEEE Transactions on Signal Processing* **2008**, *56*, 5457–5467.
27. MacKay, D.J. Information-Based Objective Functions for Active Data Selection, 1992.
28. Seo, S.; Wallat, M.; Graepel, T.; Obermayer, K. Gaussian process regression: Active data selection and test point rejection. *Advances in Neural Information Processing Systems* **2000**, *12*, 610–616.
29. Meliou, A.; Guestrin, C.; Hellerstein, J.M. Approximating sensor network queries using in-network summaries. In Proceedings of the Proceedings of the 8th International Conference on Information Processing in Sensor Networks, IPSN 2009, April 13–16, 2009, San Francisco, California, USA, 2009, pp. 229–240.
30. Kamal, A.R.M.; Hamid, M.A. Reliable data approximation in wireless sensor network. *Procedia Computer Science* **2013**, *19*, 1046–1051.

31. Kamal, A.R.M.; Razzaque, M.A.R.; Nixon, P. 2pda: two-phase data approximation in wireless sensor network. In Proceedings of the Proceedings of the 7th ACM workshop on Performance evaluation of wireless ad hoc, sensor, & ubiquitous networks. ACM, 2010, pp. 1–8.
32. Srisooksai, T.; Keamarungsi, K.; Lamsrichan, P.; Araki, K. Practical data compression in wireless sensor networks: A survey. *ICT Express* **2011**, *1*, 59–63.
33. Sheltami, T.; Musaddiq, M.; Shakshuki, E. Data compression techniques in Wireless Sensor Networks. *Future Generation Computer Systems* **2016**, *64*, 151–162.
34. Aquino, J.F.; Nakamura, E.F.; Loureiro, A.A.; Endler, M. A differential coding algorithm for wireless sensor networks. In Proceedings of the 2008 IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications. IEEE, 2008, pp. 1–5.
35. Liew, S.C.; Liew, S.W.; Zain, J.M. Reversible Medical Image Watermarking For Tamper Detection And Recovery With Run Length Encoding Compression. *World Academy of Science, Engineering and Technology* **2010**, *4*, 674–679.
36. Huffman, D.A. A Method for the Construction of Minimum-Redundancy Codes. *Proceedings of the IRE* **1952**, *40*, 1098–1101.
37. Medeiros, H.P.; Maciel, M.C.; Souza, R.D.; Pellenz, M.E. Lightweight Data Compression in Wireless Sensor Networks Using Huffman Coding. *Sensors* **2015**, *15*, 29089–29108.
38. Apostolico, A. Fast gapped variants for Lempel–Ziv–Welch compression. *Information and Computation* **2007**, *205*, 1012–1026.
39. Witten, I.H.; Neal, R.M.; Cleary, J.G. Arithmetic coding for data compression. *Communications of the ACM* **1987**, *30*, 520–540.
40. Manzini, G. An Analysis of the Burrows-Wheeler Transform. *Dipartimento di Informatica, Università del Piemonte Orientale, Italy* **2001**.
41. Chen, C.; Zhang, L.; Tiong, R.L.K. A new lossy compression algorithm for wireless sensor networks using Bayesian predictive coding. *Wireless Networks* **2020**, *26*, 5535–5547.
42. Dias, G.M.; Bellalta, B.; Oechsner, S. A Survey about Prediction-based Data Reduction in Wireless Sensor Networks. *ACM Computing Surveys (CSUR)* **2016**, *V*, A.
43. Fildes, R.; Hibon, M.; Makridakis, S.; Meade, N. Generalising about univariate forecasting methods: further empirical evidence. *International Journal of Forecasting* **1998**, *14*, 339–358.
44. Jacquot, A.; Chanet, J.P.; Hou, K.M.; Diao, X.; Li, J.J. LiveNCM: A new wireless management tool. In Proceedings of the IEEE AFRICON 2009, Nairobi, Kenya, 2009; p. 6. ZSCC: 0000007.
45. Alsheikh, M.A.; Lin, S.; Niyato, D.; Tan, H.P. Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications. *arXiv preprint arXiv:1405.4463* **2015**.
46. Enami, N.; Moghadam, R.A.; Dadashtabar, K.; Hoseini, M. Neural Network Based Energy Efficiency in Wireless Sensor Networks: A Survey. *International Journal of Computer Science & Engineering Survey* **2010**, *1*, 39–55.
47. Cheng, H.; Xie, Z.; Wu, L.; Yu, Z.; Li, R. Data prediction model in wireless sensor networks based on bidirectional LSTM. *EURASIP Journal on Wireless Communications and Networking* **2019**, *2019*, 203.
48. Mesin, L.; Aram, S.; Pasero, E. A neural data-driven algorithm for smart sampling in wireless sensor networks. *EURASIP Journal on Wireless Communications and Networking* **2014**, *2014*, 23.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.