

The Impact of Dual Prediction Schemes on the Reduction of the Number of Transmissions in Sensor Networks

Gabriel Martins Dias*, Boris Bellalta, Simon Oechsner

Pompeu Fabra University, Barcelona, Spain

Abstract

Future Internet of Things (IoT) applications will require that billions of wireless devices transmit data to the cloud frequently. However, the wireless medium access is pointed as a problem for the next generations of wireless networks; hence, the number of data transmissions in Wireless Sensor Networks (WSNs) can quickly become a bottleneck, disrupting the exponential growth in the number of interconnected devices, sensors, and amount of produced data. Therefore, keeping a low number of data transmissions is critical to incorporate new sensor nodes and measure a great variety of parameters in future generations of WSNs. Thanks to the high accuracy and low complexity of state-of-the-art forecasting algorithms, Dual Prediction Schemes (DPSs) are potential candidates to optimize the data transmissions in WSNs at the finest level because they facilitate for sensor nodes to avoid unnecessary transmissions without affecting the quality of their measurements. In this work, we present a sensor network model that uses statistical theorems to describe the expected impact of DPSs and data aggregation in WSNs. We aim to provide a foundation for future works by characterizing the theoretical gains of processing data in sensors and conditioning its transmission to the predictions' accuracy. Our simulation results show that the number of transmissions can be reduced by almost 98% in the sensor nodes with the highest workload. We also detail the impact of predicting and aggregating transmissions according to the parameters that can be observed in common scenarios, such as sensor nodes' transmission ranges, the correlation between measurements of different sensors, and the period between two consecutive measurements in a sensor.

Keywords: sensor networks, data science, predictions, data reduction, model

1. Introduction

Wireless sensor nodes (sensor nodes, for brevity) are small computer devices with a radio antenna. They are often equipped with sensors that are capable of sensing one or more environmental parameters. As an example, temperature and relative humidity sensors are some of the cheapest and smallest sensor chips

*Corresponding author

Email address: gabriel.martins@upf.edu (Gabriel Martins Dias)

available and are commonly used in real world applications.

Sensor nodes are usually organized as Wireless Sensor Networks (WSNs) that, at the application layer, have two fundamental roles: Gateways (GWs) and ordinary sensor nodes. Ordinary sensor nodes are typically close to the data origin and may just perform default sensing tasks and transmit their measurements via radio to a GW. GWs are responsible for forwarding the gathered data to WSNs' managers and for disseminating occasional instructions and updates to sensor nodes.

The WSNs' growth can be described by the increasing number of wireless sensor nodes measuring and reporting data to GWs, and by the diversity of data types transmitted in WSNs. As a consequence of this growth, modern applications of WSNs do not simply monitor changes in the environment anymore; they also trigger reactions to these changes. For example, in agriculture, sensed data could be used to apply pesticides after detecting that the number of insects exceeded a certain threshold [1]. Such a threshold, in turn, may vary according to the season or get affected by temperature and relative humidity changes during the days. In these applications, sensor nodes may need a high number of transmissions to communicate the number of insects, temperature, and relative humidity. If the WSN cannot handle all transmissions that sensor nodes make, it will collapse and end up in significant economic losses.

In other cases, losses can be even worse. For example, structural health monitoring for aircraft can include engine control systems that rely on enhanced data analysis to detect accidents and report unexplained phenomena to people responsible for maintaining their safety and healthy conditions [2]. Extended data analysis usually require, in comparison to the simplest monitoring tasks, more parameters and a higher amount of informative data. Therefore, in these cases, more sensor nodes transmitting higher amounts of data might collapse the WSN and mispredict accidents, resulting in living losses.

These situations help us to understand that WSNs are data-oriented networks, i.e., the data produced by the sensor nodes is their most valuable asset [3]. However, sensor nodes typically have limited energy resources, and transmitting data is the task that drains the most battery power. Hence, several works addressed sensor nodes' energy management as one of the biggest challenges for WSN applications [4].

Fortunately, new technologies can harvest energy from solar, mechanical and thermal energy sources [5, 6], showing that this problem can be—at least, partially—overcome in the next years. However, in addition to the energy consumption of the sensor nodes, the efficient use of spectrum resources has been pointed out as one of the key challenges that will affect the next generation of wireless networks, for instance, WLANs, 4G and 5G networks [7, 8, 9].

In this work, we present a WSN model for data transmissions in monitoring applications. As future generations of sensor nodes tend to evolve and eventually incorporate other Internet of Things (IoT) devices (such as smartphones, home appliances, and vehicles), variables as sensor nodes' connectivity and their maximum distance to the GW are considered in the model. In the end, focusing on the challenges for the

next generations of WSNs, we use statistical theorems to draw conclusions about how efficient is to adopt a Dual Prediction Scheme (DPS) to reduce the number of data transmissions in a monitoring application. As an outcome, one of the main contributions of this work is a mathematical model that formulates the number of transmissions in a WSN that simultaneously adopts a DPS and aggregation to reduce its number of transmissions.

Our goal is to provide a sufficiently strong background on which future applications can rely to create smarter and more complex systems. In the next sections, we will evaluate the gains that can be obtained by reducing the number of data transmissions and compare with the costs of choosing a certain prediction model, considering its expected (in)accuracy and the correlation between the measurements made by different sensor nodes.

The rest of the paper is organized as follows: Section 2 describes how DPSs can be adopted in WSNs, based on previous studies in this field; Section 3 describes our WSN transmission model in detail; Section 4 describes how we use our model to represent a DPS and estimate its benefits in WSNs; Section 5 models the number of transmissions and the energy consumption expected when adopting the data aggregation or the data prediction methods; Section 7 shows the current state of the art, giving a wider perspective and highlighting the contributions of this work; and Section 8 shows our conclusions and outlines for future work.

2. Background - Dual Prediction Schemes

The term *prediction* can either refer to the process of inferring missing values in a dataset based on statistics or empirical probability, or to the estimation of future values based on the historical data. A *prediction method* P is a deterministic algorithm that produces predictions based on two input variables: a set of *observed values* X and a *set of parameters* θ . A *prediction model* p is an instance of a prediction method P , such that $p_\theta(X) = P(X, \theta)$. Thus, a prediction model can be summarized as P and θ .

The values of θ can be provided by a utility function that measures the predictions' accuracy, models' complexity or information loss. Thus, *choosing a prediction model* means finding the values of θ that best summarizes the current measurements under the criteria adopted by the utility function (e.g., the minimum information loss estimated).

To choose a prediction model in a DPS, sensor nodes report all the data that they have measured to the GWs during an *initialization phase* [10]. The choice of the prediction model may be made independently in GW and sensor nodes; alternatively, each sensor node can choose its prediction model and transmits the parameters to the GW; or, finally, the GW may be given the autonomy to choose prediction models for all sensor nodes. The decision about where to choose the prediction model can be taken at runtime, and it is not necessary to fix a strategy for the whole WSN's lifetime [11].

After choosing (and occasionally sharing) the prediction models, sensor nodes can exploit their proximity

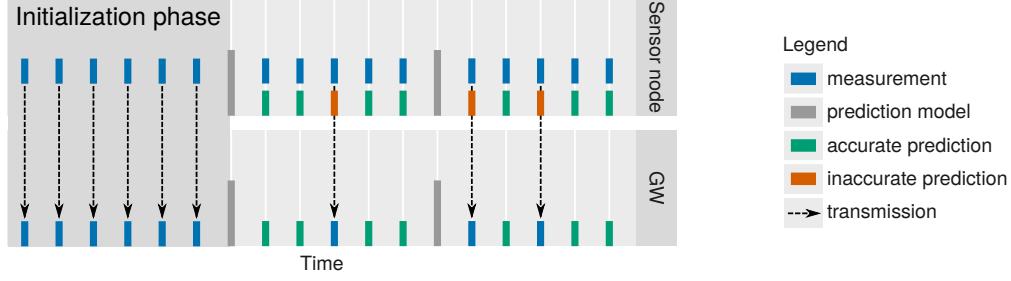


Figure 1: A variant of the DPS with independent model generation. The sensor node and the GW can compute the same prediction model because they are programmed to use the same data.

to the data's origin. That is, they can compare predicted values with real measurements and transmit the actual measurements only if the predictions are inaccurate. From time to time, if predictions' accuracy gets compromised, sensor nodes and GW may begin a new *initialization phase* and choose new prediction models.

2.1. Independent model choice

The *initialization phase* ensures that the GW will have complete information about the environment before any prediction model is chosen. Therefore, after this phase, the GW can choose the same prediction models as sensor nodes without making any new transmission. Figure 1 illustrates the sensor nodes' and GW's behaviors. New prediction models can be regularly chosen based on the knowledge simultaneously available to sensor nodes and GW. As a drawback, the variety of the prediction models is restricted by the memory and computing power limitations of sensor nodes.

The Least Mean Squares (LMS) method has provided accurate predictions in simulations where sensor nodes and GW generated their prediction models independently [10, 12, 13]. For instance, in a particular scenario, only 10% of the measurements would be necessary to monitor room temperature accurately [10].

2.2. Model choice in sensor nodes

Alternatively, prediction models can be chosen in sensor nodes, as illustrated in Figure 2. As before, sensor nodes start transmitting all the measurements to the GW. However, a new responsibility is assigned to sensor nodes: after collecting local measurements and choose a prediction model that fits the current environment, they must communicate the prediction model to the GW. The main advantage of this mechanism is that sensor nodes can decide for new prediction models using all the measured data, instead of using only the information that they share with the GW. On the other hand, sensor nodes need extra transmissions to inform the GW about their decisions.

Simulation results using real data from WSNs showed that this approach could reduce the number of data transmissions using AutoRegressive (AR), AutoRegressive Integrated Moving Average (ARIMA), or Exponential Smoothing (ES) models with neither exceeding the constrained memory nor the computational

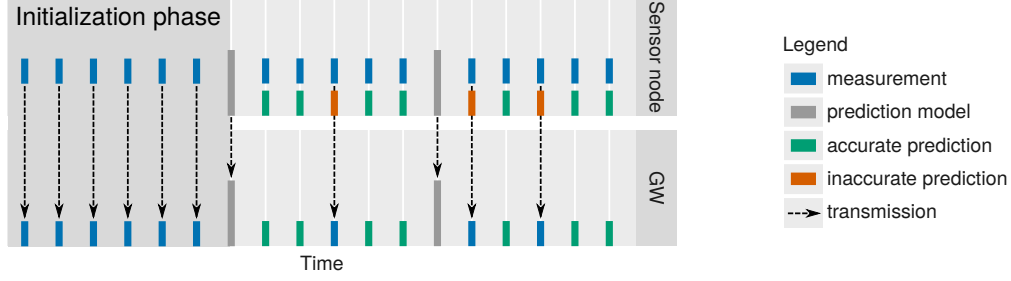


Figure 2: As sensor nodes can overhear their neighbors' data without overloading the network or congesting the medium, they may locally decide the best prediction method and later inform their decision to the GWs.

resources of typical wireless sensor nodes [14, 15, 16]. Alternatively, a hybrid mechanism may improve the quality of the predictions if sensor nodes have the autonomy to adopt more complex prediction methods (e.g., Artificial Neural Networks (ANNs)) when the simplest predictions (e.g., ARIMA) are inaccurate [17]. Only in the worst case, if the difference between the measurements and the predictions using the most complex method also exceed the acceptance threshold, sensor nodes are responsible for transmitting the real measurements to the GW.

More recently, in [18, 19], the authors gave to the sensor nodes the ability to take decisions locally using Gaussian Processes and Stochastic Gradient Descent regression, which require much higher computational power than the traditional methods. Following the trend of adopting complex prediction methods in sensor networks, in [20], the authors incorporated information theory in their analysis and described a method that can accurately predict and evolve their prediction models.

2.3. Model choice in the Gateway

In this type of DPS, the GW is responsible for periodically choosing and transmitting new prediction models' parameters and error acceptance levels to sensor nodes, as shown in Figure 3. Generating the prediction models in the GW exploits the asymmetric computational power availability in WSNs: GWs usually have cheaper energy sources and more resources (such as memory and processing power) than ordinary sensor nodes that mainly measure and report environmental data [21, 22]. Eventually, GWs can rely on cloud services to analyze the collected data and choose more accurate prediction models [23]. For example, ANNs can provide higher accuracy than other methods, but they may not fit sensor nodes' constraints because building an ANN requires a computation intensive training phase over a large amount of data.

Additionally, the GW can assume the responsibility of adapting sensor nodes' operations according to the potential savings that predictions may introduce. In such cases, the GW can estimate if it is worth to make predictions in sensor nodes, based on the relation between the predictions' accuracy, the correlation between measurements, and the error tolerated by the user [24]. According to the expected gains, sensor nodes can be set to: (i) go to sleep mode without making any measurement; (ii) make measurements and transmit every

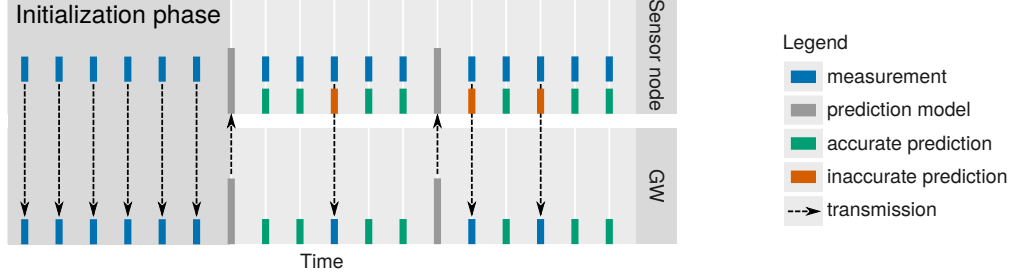


Figure 3: In a DPS, a measurement is only transmitted if its forecast is inaccurate. The GWs may be responsible for transmitting new prediction models every time interval after the initialization phase.

measurement done; or (iii) make measurements, transmit them to the GW whenever the prediction differs by more than an acceptance value.

3. A WSN transmission model

Langendoen and Meier [25] presented a ring model for WSN topologies to describe a multi-hop network based on the average number of neighbors (C) of a sensor node and the number of hops from the GW to the furthest nodes (D). Assuming a uniform node density on the plane and defining it as $C + 1$ nodes per unit disk, the first ring ($d = 1$) is expected to have C nodes. Figure 4 shows an example of a sensor network based on this model with $D = 3$ and $C = 5$.

In this model, the GW is always in ring zero, and transmissions made by a component (either the GW or a sensor node) can reach neighbors that are up to one unit of length from it. The set of neighbors of a sensor node i is defined by all sensor nodes in the unit disk centered in i . The unit disk represents the sensor node's transmission range and does not necessarily imply that neighbor sensor nodes will establish a direct communication at the routing layer.

In fact, communication links are defined by underlying routing protocols. Langendoen and Meier assumed that these protocols aim to keep the smallest number of hops in a WSN and that sensor nodes only transmit to sensor nodes in the previous ring, i.e., the next ring closer to the GW. For example, to reach the GW from ring d , we can expect a d -hop transmission. Therefore, the distance from the GW also defines in which ring a sensor node is placed.

The expected number of sensor nodes N_d in ring d can be calculated based on the surface area of the annulus¹:

¹The region bounded by two concentric circles.

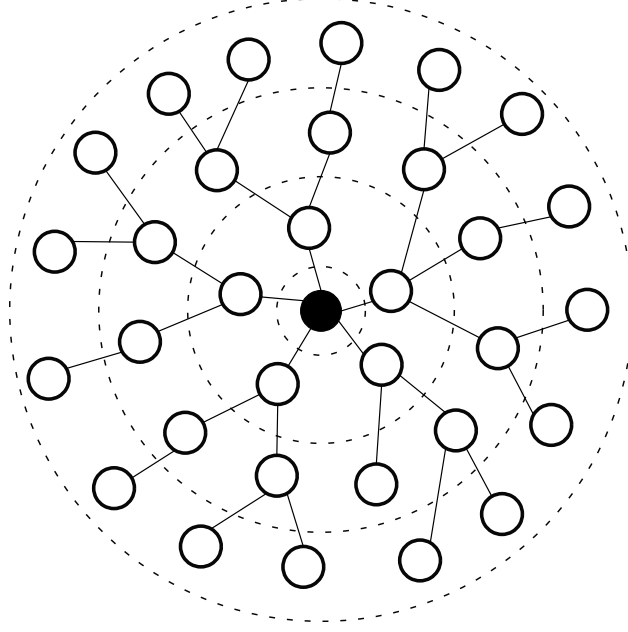


Figure 4: Sensor network model based on the density of the sensor nodes and their coverage. Each node has an average of five ($C = 5$) neighbors at the physical layer, and the vertices represent communication links established in an average (optimistic) scenario. The dark circle represents the GW.

$$N_d = \begin{cases} 0, & \text{if } d = 0 \\ Cd^2 - C(d-1)^2 = (2d-1)C, & \text{otherwise} \end{cases} \quad (1)$$

The number of nodes in the WSN is equal to CD^2 . Given that the first ring has C sensor nodes, it is expected C branches starting in the GW with D^2 sensor nodes each. In this work, each of these branches will be referenced as a *sub-tree*.

Assuming a sensor node in ring d , the expected number of direct children (I_d) can be calculated by the relation N_{d+1}/N_d . This value does not depend on the value of C :

$$I_d = \begin{cases} 0, & \text{if } d = D \\ \frac{2d+1}{2d-1}, & \text{otherwise} \end{cases} \quad (2)$$

3.1. Model extension

We assume that the number of transmissions and receptions made by sensor nodes is the primary concern in monitoring applications, not only due to the challenge to access the medium but also due to the energy required for the external communication. Although these challenges are commonly observed in irregular real-world topologies, they are often neglected by other models that do not give special attention to the sensor nodes that may have the highest workloads [26]. Having said that, we highlight that the main advantage of

Parameter	Description
f	Number of measurements per time slot
T	Period between the choice of two new prediction models
C	Expected number of neighbors of each sensor node
D	Expected number of rings/hops in the sensor network
ρ	Correlation between measurements in a <i>sub-tree</i>
α	Expected prediction's accuracy

Table 1: Parameters taken into account to calculate the number of transmissions and receptions using the model.

this model is its simplicity to identify and describe the operation of the bottlenecks in a sensor network, i.e., the sensor nodes in the first ring.

In this work, we will extend the original model and derive the number of transmissions based on the sensor nodes' positions. First, we define the set of children nodes of a sensor node i in ring d as $H_{i,d}$. We define the expected cardinality of $H_{i,d}$ as K_d . The value of K_d is the number of direct children times the expected number of their children plus one (representing themselves):

$$K_d \triangleq |H_{i,d}| = \begin{cases} 0, & \text{if } d = D \\ I_d(K_{d+1} + 1), & \text{otherwise.} \end{cases} \quad (3)$$

Recall that the expected number of sensor nodes is CD^2 , and the first ring is expected to have C nodes. Thus, the expected number of children of the sensor nodes in the first ring (i.e., K_1) is $D^2 - 1$ if $D > 1$.

3.1.1. Node-to-GW transmissions

In monitoring applications, sensor nodes usually measure and transmit their data at a pre-defined time interval ($1/f$) that can vary from few seconds to hours. The number of measurements per second (f), the period between choosing new prediction models (T) and the other parameters shown in Table 1 may vary from case to case.

In the simplest approach, measurements are transmitted right after their creation. We will assume this behavior as the baseline for further comparisons. Alternatively, these transmissions, which we call *node-to-GW*, may not happen right after measurements' creation if sensor nodes aggregate the data received from other sensor nodes or past measurements. The impact of aggregating transmissions will also be modeled in the following.

Given that sensor nodes must forward data from their children towards the GW, the expected number of transmissions ($E(S_d)$) during a period of $1/f$ seconds in a sensor node i in ring d is

$$E(S_{i,d}) = (K_d + 1), \quad (4)$$

and the number of receptions is

$$E(R_{i,d}) = K_d. \quad (5)$$

Finally, the total number of transmissions during a period T in a sensor node i is the sum of transmissions and receptions:

$$\begin{aligned} E(X_{i,d}) &= E(S_d) + E(R_d) \\ &= ((K_d + 1) + K_d) fT \\ &= (2K_d + 1) fT. \end{aligned} \quad (6)$$

Based on (3), we can affirm that $K_1 > K_d$, if $d > 1$. Applying this inequality to (4), (5), and (6), we mathematically show that, if $d > 1$, then $E(S_{i,1}) > E(S_{i,d})$, $E(R_{i,1}) > E(R_{i,d})$, and $X_{i,1} > X_{i,d}$. Thus, in a homogeneous sensor network, sensor nodes in the first ring make more transmissions and are the bottlenecks that limit the number of transmissions in their *sub-trees*, according to their capacity. It shows the importance of focusing on optimizing the number of transmissions in the first ring because it can prevent new sensor nodes of being appended to the network.

3.1.2. GW-to-node transmissions

GW-to-node transmissions are those initiated by the GW, for example, to change a configuration or update the software in the sensor nodes. Assuming one unicast transmission per packet, if the GW transmits a packet to every sensor node in the WSN, a sensor node i in ring d will receive one *GW-to-node* transmission to itself, plus K_d transmissions (one per child), which must be forwarded towards their receivers. Therefore, the number of transmissions and receptions at a sensor node i in ring d is

$$E(S_{i,d}^*) = K_d, \quad (7)$$

and

$$E(R_{i,d}^*) = K_d + 1. \quad (8)$$

In this case, the average number of transmissions made by the GW to a *sub-tree* is D^2 , i.e., the number of nodes in each *sub-tree*. Therefore, a sensor node in the first ring will make $K_1 + 1 = D^2$ receptions and $K_1 = D^2 - 1$ transmissions.

4. Modeling Dual Prediction Schemes

As explained in Section 2, DPSs exploit the proximity of the sensor nodes to the sources of the data, avoiding unnecessary transmissions and handling occasional sensor nodes' hardware limitations that might reduce WSNs' lifetime. A DPS has two tasks that may be executed either by GWs or by sensor nodes, namely

the *prediction model choice* and the *prediction model dissemination*. The *dissemination* is the process of transmitting the prediction model either from sensor nodes to the GW or from the GW to sensor nodes.

In the following, we describe the impact of these tasks in the network load, concerning the number of transmissions. Before that, we discuss the assumptions and limitations of this model.

4.1. Assumptions and limitations

In this model, we assume that the quality of the measurements provided by a WSN can be scaled as “acceptable” if the values at the GW do not differ by more than a certain threshold. Since sensor nodes can compare their predictions with real measurements locally (without making any transmission), no transmission will be required if a prediction is accurate, i.e., it does not differ by more than an acceptance threshold from the measured value.

In some cases, WSNs’ managers have no information about the statistics of the data that is going to be retrieved by the sensor nodes. Thus, it may be necessary a long *initialization phase* before starting to make predictions. For example, schemes that use advanced prediction methods, like ANNs, require larger amounts of data to find stable models, due to their high complexity and the vast number of parameters to estimate [27]. We do not include this phase in this model because we assume that the GW has no energy constraints.

In this work, we do not expect distributed algorithms, i.e., sensor nodes do not have to synchronize with their neighbors. However, this can be easily extended from our model, given the number of expected neighbors of each sensor node.

Moreover, object tracking and event detection applications are not under the scope of this work because they usually have other requirements than those we assume in this model, such as higher reliability or lower delays [28].

Finally, our model is designed to represent an average WSN with connectivity between its sensor nodes at the application layer. Faulty sensors may affect the sensor node distribution, their density and force a new routing strategy. Therefore, small WSNs with low sensor nodes’ density would require special attention, as the number of children of a sensor node may change and eventually overload it with extra transmissions. On the other hand, large and dense WSNs may not be impacted by small numbers of faulty sensor nodes, and our model can still be valid.

4.2. Prediction model choice and dissemination

In a DPS, the same prediction model is shared between a sensor node and the GW. Each sensor node (or group of sensor nodes) has its prediction model, and the prediction models in a WSN can be independently chosen by both (sensor nodes and GW) without making any new transmission. Alternatively, prediction models can be chosen in the GW or in sensor nodes. In case that prediction models are chosen in sensor

nodes, the GW must receive the parameter values and, in some cases, also the prediction method selected. On the other hand, if the GW is responsible for choosing prediction models, sensor nodes must be informed about the decisions taken.

Assuming that the *dissemination* of a prediction model is made through a unicast transmission, sensor nodes in the first ring will receive and forward every transmission to their children towards the proper destinations. Thus, if the GW is responsible for generating the prediction models, the sensor nodes in the first ring will have to forward the transmissions to their children. In such cases, a sensor node in the first ring will receive D^2 packets. From these packets, $D^2 - 1$ will be forwarded to its children. Therefore, to disseminate the prediction models generated in the GW, there will be

$$\begin{aligned} E(X_{\text{DIS-GW}}) &= E(R_{i,1}^*) + E(S_{i,1}^*) \\ &= D^2 + (D^2 - 1) \\ &= 2D^2 - 1 \end{aligned} \tag{9}$$

transmissions (summing transmissions and receptions) in each sensor node in the first ring.

Analogously, in case that prediction models are chosen in the sensor nodes, every sensor node in the first ring will make D^2 transmissions to the GW after receiving $D^2 - 1$ prediction models. Thus, the number of transmissions at the first ring will be, once again, equal to $2D^2 - 1$.

If packets to the same *sub-tree* are aggregated, sensor nodes in the first ring will receive only one packet that will be split before being retransmitted to their direct children in the second ring. In such cases, a sensor node in the first ring will need

$$E(X_{\text{DIS-GW-AGG}}) = E(X_{\text{DIS-SN-AGG}}) = I_1 + 1 \tag{10}$$

transmissions to *disseminate* the prediction models, where, from (2), $I_1 = 3$, if $D > 1$.

Finally, if the GW uses broadcast (or multicast) transmissions, sensor nodes will receive and forward only one packet, i.e.,

$$E(X_{\text{DIS-GW-BC}}) = 1. \tag{11}$$

4.3. Impact of predictions in the number of transmissions

As described before, adopting a data prediction scheme can benefit the WSN reducing the number of transmissions and optimizing the medium access control, which may eventually reduce energy consumption and extend the WSN's lifetime. To estimate the number of transmissions in homogeneous networks, we develop a formula based on the predictions' accuracy and the correlation of the monitored data.

Let us assume that α_i is the accuracy of the predictions in sensor node i , i.e., α_i is the probability that a measurement of i matches to the prediction and does not have to be transmitted to the GW, and $\alpha_i^c = 1 - \alpha_i$.

Therefore, the expected number of transmissions and receptions in a sensor node i during a time interval of $1/f$ seconds (i.e., between two measurements) is respectively represented by $E(S'_{i,d})$ and $E(R'_{i,d})$ as

$$E(S'_{i,d}) = \alpha_i^c + \sum_{j \in H_{i,d}} \alpha_j^c, \quad (12)$$

and

$$E(R'_{i,d}) = \sum_{j \in H_{i,d}} \alpha_j^c. \quad (13)$$

Considering an eventual *dissemination* of the prediction models, the expected number of transmissions and receptions during a period of T seconds is

$$\begin{aligned} E(X'_{i,d}) &= (E(S'_{i,d}) + E(R'_{i,d}))fT + E(X_{\text{DIS}}) \\ &= (\alpha_i^c + \sum_{j \in H_{i,d}} \alpha_j^c + \sum_{j \in H_{i,d}} \alpha_i^c)fT + E(X_{\text{DIS}}) \\ &= (\alpha_i^c + 2 \sum_{j \in H_{i,d}} \alpha_j^c)fT + E(X_{\text{DIS}}) \end{aligned} \quad (14)$$

Notice that a low accuracy in predictions used in sensor nodes that are far from the GW has a higher impact on the total number of WSN transmissions than a low accuracy in predictions used in sensor nodes in the first rings. However, concerning the number of transmissions at a single sensor node, the bottleneck of the WSN is still represented by sensor nodes in the first ring.

Let us define the minimum average accuracy (α_{\min}) necessary to reduce the number of transmissions, according to the size of the network and its number of rings. This value can be used to define the maximum number of transmissions ($E(\max(S'_{i,d}))$) and receptions ($E(\max(R'_{i,d}))$) in a sensor node i in ring d :

$$\begin{aligned} E(\max(S'_{i,d})) &= (1 - \alpha_{\min}) + \sum_{j \in H_{i,d}} (1 - \alpha_{\min}) \\ &= (1 + K_d) (1 - \alpha_{\min}) \end{aligned} \quad (15)$$

and

$$\begin{aligned} E(\max(R'_{i,d})) &= \sum_{j \in H_{i,d}} (1 - \alpha_{\min}) \\ &= K_d (1 - \alpha_{\min}) \end{aligned} \quad (16)$$

Recall that $K_d \triangleq |H_{i,d}|$, for a sensor node i in ring d . Therefore,

$$E(X'_{i,d}) \leq ((1 + K_d) (1 - \alpha_{\min}) + K_d (1 - \alpha_{\min}))fT + E(X_{\text{DIS}}) \quad (17)$$

Finally, the use of predictions will reduce the number of transmissions if $X'_{i,d} < X_{i,d}$. After some mathematical development shown in Appendix A, we arrive at the following expression for the minimum average accuracy of the predictions:

$$\alpha_{\min} > \frac{E(X_{\text{DIS}})}{(2D^2 - 1)fT} \quad (18)$$

In conclusion, if prediction models are not independently generated in GWs and sensor nodes, the DPS requires a minimum accuracy to ensure the reduction in the number of transmissions. Hence, the minimum accuracy is a lower bound that depends only on the network layout (i.e., the number of rings D), the frequency of the measurements (f) and the time between choosing two prediction models (T). If the predictions' accuracy does not reach this limit, there will exist two actions to improve the network operation, either to set new values for f and T or to turn the DPS off.

4.4. Impact of predictions in the number of transmissions

As described before, adopting a data prediction scheme can benefit the WSN reducing the number of transmissions and optimizing the medium access control, which may eventually reduce energy consumption and extend the WSN's lifetime. To estimate the number of transmissions in homogeneous networks, we develop a formula based on the predictions' accuracy and the correlation of the monitored data.

Let us assume that α_i is the accuracy of the predictions in sensor node i , i.e., α_i is the probability that a measurement of i matches to the prediction and does not have to be transmitted to the GW, and $\alpha_i^c = 1 - \alpha_i$. Therefore, the expected number of transmissions and receptions in a sensor node i during a time interval of $1/f$ seconds (i.e., between two measurements) is respectively represented by $E(S'_{i,d})$ and $E(R'_{i,d})$ as

$$E(S'_{i,d}) = \alpha_i^c + \sum_{j \in H_{i,d}} \alpha_j^c, \quad (19)$$

and

$$E(R'_{i,d}) = \sum_{j \in H_{i,d}} \alpha_j^c. \quad (20)$$

Considering an eventual *dissemination* of the prediction models, the expected number of transmissions and receptions during a period of T seconds is

$$\begin{aligned} E(X'_{i,d}) &= (E(S'_{i,d}) + E(R'_{i,d}))fT + E(X_{\text{DIS}}) \\ &= (\alpha_i^c + \sum_{j \in H_{i,d}} \alpha_j^c + \sum_{j \in H_{i,d}} \alpha_j^c)fT + E(X_{\text{DIS}}) \\ &= (\alpha_i^c + 2 \sum_{j \in H_{i,d}} \alpha_j^c)fT + E(X_{\text{DIS}}) \end{aligned} \quad (21)$$

Note that a low accuracy in predictions used in sensor nodes that are far from the GW has a higher impact on the total number of WSN transmissions than a low accuracy in predictions used in sensor nodes in the first rings. However, concerning the number of transmissions at a single sensor node, the bottleneck of the WSN is still represented by the sensor nodes in the first ring.

Let us define the minimum average accuracy (α_{\min}) necessary to reduce the number of transmissions, according to the size of the network and its number of rings. This value can be used to define the maximum number of transmissions ($E(\max(S'_{i,d}))$) and receptions ($E(\max(R'_{i,d}))$) in a sensor node i in ring d :

$$\begin{aligned} E(\max(S'_{i,d})) &= (1 - \alpha_{\min}) + \sum_{j \in H_{i,d}} (1 - \alpha_{\min}) \\ &= (1 + K_d) (1 - \alpha_{\min}) \end{aligned} \quad (22)$$

and

$$\begin{aligned} E(\max(R'_{i,d})) &= \sum_{j \in H_{i,d}} (1 - \alpha_{\min}) \\ &= K_d (1 - \alpha_{\min}) \end{aligned} \quad (23)$$

Recall that $K_d \triangleq |H_{i,d}|$, for a sensor node i in ring d . Therefore,

$$E(X'_{i,d}) \leq ((1 + K_d) (1 - \alpha_{\min}) + K_d (1 - \alpha_{\min}))fT + E(X_{\text{DIS}}) \quad (24)$$

Finally, the use of predictions will reduce the number of transmissions if $X'_{i,d} < X_{i,d}$. After some mathematical development shown in Appendix A, we arrive at the following expression for the minimum average accuracy of the predictions:

$$\alpha_{\min} > \frac{E(X_{\text{DIS}})}{(2D^2 - 1)fT} \quad (25)$$

In conclusion, intuitively, if prediction models are independently generated in GWs and sensor nodes, the DPS does not require a minimum accuracy to ensure the reduction in the number of transmissions. Otherwise, if prediction models are not independently generated in GWs and sensor nodes, the extra transmissions used to disseminate new prediction models may turn the prediction scheme into an inefficient option. Hence, the efficiency of a DPS also depends on how many transmissions are required for disseminating the prediction models because the number of transmissions will be proportional to the number of hops between sensor nodes and the GW.

Moreover, the minimum accuracy is a lower bound that depends only on the network layout (i.e., the number of rings D), the frequency of the measurements (f) and the time between two predictions (T). Therefore, if the predictions' accuracy does not reach this limit, there will exist three actions to improve the

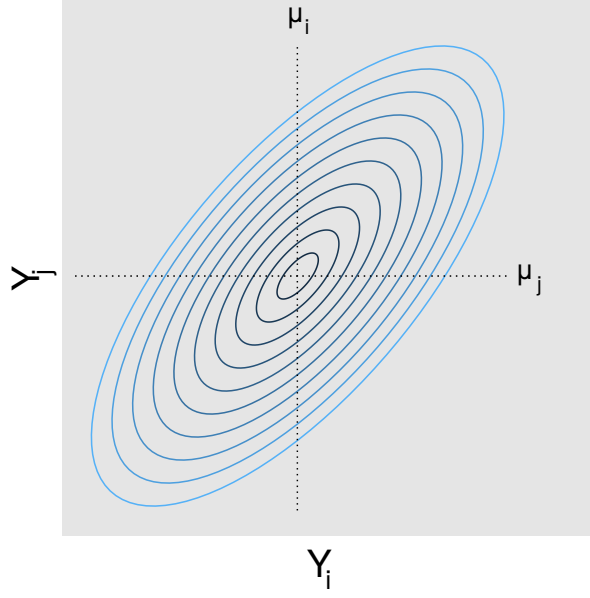


Figure 5: Values of Y_i and Y_j are correlated ($\rho_{i,j} = 0.7$), and each line represents a different density of points.

network operation, either to set new values for f and T , to adopt a DPS with independent model generation, or to turn the DPS off.

4.5. Impact of predictions and data aggregation

Additionally to DPSs, it may be possible to adopt aggregation schemes in sensor nodes, such that a sensor node aggregates data received from its children and transmits only after making its measurement. In the following, we model an aggregation scheme and compare its efficiency with the use of DPSs. Finally, we estimate the reduction in the number of transmissions if both techniques are simultaneously adopted.

To make it clear for the reader, we introduce a scenario with only two sensor nodes to clarify the normalization of the data and its application. Later, we will extend the model to a more complex scenario with D rings.

4.5.1. Network with two sensor nodes

Let us consider a section of the sensor network with the GW and a sensor node i with a single child j . Due to the sensor network's topology, transmissions from sensor node j can reach the GW only through i . Thus, every $1/f$ seconds, i may transmit to the GW if its prediction has failed or if it had happened to j .

We assume that the measurements of i and j follow the Normal distributions respectively represented by $Y_i = N(\mu_i, \sigma_i^2)$ and $Y_j = N(\mu_j, \sigma_j^2)$. Such a Multivariate Normal (MVN) distribution can be defined based on the correlation between their values, i.e., the relationship between each pair of measurements made by i and j . An illustration of the MVN density containing Y_i and Y_j is shown in Figure 5.

Assuming that the predictions (\bar{y}) are not biased (i.e., $\bar{y} = \mu$), we may also approximate them to Normal distributions² and label an outcome as incorrect whenever a measurement falls outside the interval defined by the acceptance threshold ε . In such cases, the probability that the sensor node j will transmit (including its measurement) after $1/f$ seconds is $1 - \alpha_j$. Hence, the probability of i receiving a packet is also $1 - \alpha_j$.

Similarly, i will transmit if the prediction about its measurement fails (i.e., it differs from the actual measurement more than the acceptance threshold ε_i) or if the prediction in sensor node j had failed and a measurement has been received. In other words, there will be a transmission if at least one of the two predictions fail.

If i can aggregate transmissions, its total number of transmissions is not a simple sum as in the case without aggregation because it depends on the correlation of the measurements of i and j . Let us assume that the correlation between Y_i and Y_j is defined by the Pearson correlation coefficient and represented by $\rho_{i,j}$. Therefore, to model the probability of having at least one wrong prediction, we must calculate the correlation matrix (Σ), which is defined as

$$\Sigma = \begin{bmatrix} \sigma_i^2 & \rho_{i,j} \sigma_i \sigma_j \\ \rho_{i,j} \sigma_i \sigma_j & \sigma_j^2 \end{bmatrix} \quad (26)$$

Finally, given the lower limits

$$l_i = \bar{y}_i - \varepsilon_i \text{ and } l_j = \bar{y}_j - \varepsilon_j, \quad (27)$$

the upper limits

$$u_i = \bar{y}_i + \varepsilon_i \text{ and } u_j = \bar{y}_j + \varepsilon_j, \quad (28)$$

and the correlation matrix (Σ), it is possible to calculate the following MVN probability:

$$F(y_i, y_j) = \frac{1}{\sqrt{|\Sigma|(2\pi)^2}} \int_{l_i}^{u_i} \int_{l_j}^{u_j} e^{(-\frac{1}{2}\theta^t \Sigma^{-1} \theta)} d\theta \quad (29)$$

The value of $F(y_i, y_j)$ represents the probability that both predictions (in i and j) are correct and can be illustrated by the density inside the crosshatched rectangle in Figure 6. Thus, the probability of at least one prediction failing can be calculated as $(1 - F(y_i, y_j))$, which, in fact, is the probability of sensor node i making a transmission after $1/f$ seconds.

Finally, considering occasional extra transmissions to disseminate the prediction model, the number of transmissions at j during a period of T seconds can be calculated as $(1 - \alpha_j)fT + E(X_{\text{DIS}})$, and the

²In Appendix B, we detail how to estimate the predictions' accuracy for normally distributed measurements, based on the user's acceptance threshold for errors.

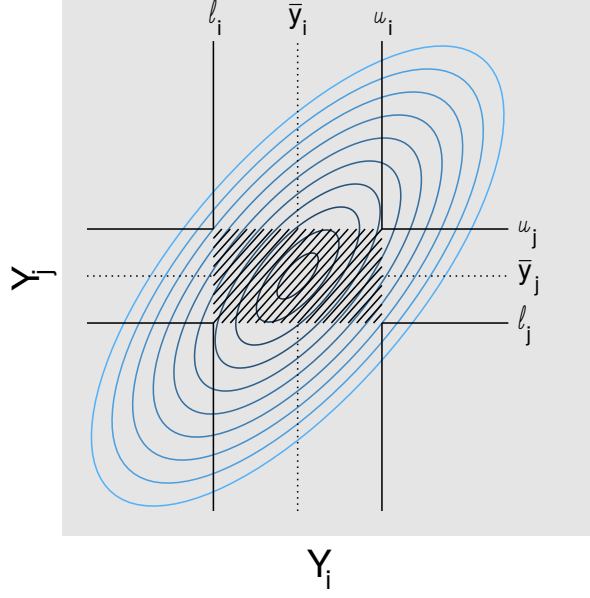


Figure 6: The hashed rectangle in the center illustrates the points in which both predictions (\bar{y}_i and \bar{y}_j) are correct.

expected sum of transmissions and receptions at sensor node i during a period of T seconds can be modeled as $((1 - F(y_i, y_j)) + (1 - \alpha_j))fT + E(X_{\text{DIS}})$.

4.5.2. Larger networks

Now, we will extend the previous example to larger sensor networks. The correlation matrix (Σ) of several data distributions can be calculated as

$$\Sigma = \begin{bmatrix} \sigma_a^2 & \rho_{a,b} \sigma_a \sigma_b & \cdots & \rho_{a,z} \sigma_a \sigma_z \\ \rho_{b,a} \sigma_b \sigma_a & \sigma_b^2 & \cdots & \rho_{b,z} \sigma_b \sigma_z \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{z,a} \sigma_z \sigma_a & \rho_{z,b} \sigma_z \sigma_b & \cdots & \sigma_z^2 \end{bmatrix}, \quad (30)$$

and, similarly to the two-dimensional model, the expected number of transmissions made by sensor node i in ring d (represented by $E(S''_{i,d})$) depends not only on its predictions but also on the predictions used in all of its children. The value of $E(S''_{i,d})$ can be calculated as

$$E(S''_{i,d}) = 1 - F(i, a, b, \dots, z), \quad (31)$$

where $\{a, b, \dots, z \in H_{i,d}\}$, and the function F is the MVN probability function integrated from the lower accepted limits to the upper accepted limits over the $k = 1 + K_d$ distributions:

$$F(i, a, b, \dots, z) = \frac{1}{\sqrt{|\Sigma|(2\pi)^k}} \int_{l_i}^{u_i} \int_{l_a}^{u_a} \int_{l_b}^{u_b} \dots \int_{l_z}^{u_z} e^{(-\frac{1}{2}\theta^t \Sigma^{-1} \theta)} d\theta, \quad (32)$$

which can be efficiently calculated with the use of Monte Carlo methods for higher dimensions [29].

The number of receptions at i ($E(R''_{i,d})$) is slightly different from the previous example, since now the sensor node may have several children in the next ring, and their transmissions happen independently. Let us define $H'_{i,d}$ as the set of direct children of i . Thus, $|H'_{i,d}| \triangleq I_d$. The expected number of receptions can be calculated as

$$E(R''_{i,d}) = \sum_{j \in H'_{i,d}} E(S''_{j,d+1}), \quad (33)$$

and the total number of transmissions and receptions is expected to be

$$E(X''_{i,d}) = (E(S''_{i,d}) + E(R''_{i,d}))f T + E(X_{\text{DIS}}). \quad (34)$$

Even though the function F has no closed formula, it is possible to set a lower bound based on a case when there is absolutely no correlation between the values measured by i and its children. When the correlation is equal to zero, the expected number of transmissions and receptions at sensor node i are the maximum possible. Considering that there will exist a transmission if at least one prediction fails, the probability of having no transmissions at i is α^{1+K_d} . Thus,

$$E(\max(S''_{i,d})) = 1 - \alpha^{1+K_d}. \quad (35)$$

Recall that i is expected to have I_d direct children and each child be part of a *sub-tree* with K_d/I_d sensor nodes. There may exist I_d independent receptions at i , and each reception may not occur with probability α^{K_d/I_d} . Thus,

$$E(\max(R''_{i,d})) = I_d (1 - \alpha^{K_d/I_d}). \quad (36)$$

Therefore,

$$E(X''_{i,d}) \leq [(1 - \alpha^{1+K_d}) + I_d (1 - \alpha^{K_d/I_d})] f T + E(X_{\text{DIS}}). \quad (37)$$

We claim that $E(X''_{i,d}) \leq E(X'_{i,d})$, which means that a mechanism that aggregates the data will not make more transmissions than the one that only makes predictions. Comparing (37) with (24), we have that for any $\alpha \in [0, 1]$ and $K_d \geq 0$, it can be shown³ that $(1 - \alpha^{1+K_d}) \leq ((1 + K_d)(1 - \alpha))$ and, hence, $E(\max(S''_{i,d})) \leq E(\max(S'_{i,d}))$. Moreover, $E(\max(R''_{i,d})) \leq E(\max(R'_{i,d}))$ and $I_d (1 - \alpha^{K_d/I_d}) \leq K_d (1 - \alpha)$,

³Based on the proof detailed in Appendix C.

which can be similarly proved to be true, since $(K_d/I_d) \geq 1$ when $K_d > 0$ and $\alpha \in [0, 1]$. In the case of being in the last ring, since there are no children ($K_d = I_d = 0$), no reception is made.

5. Model experimentation

Using the model presented before, we can estimate the effects of adopting a prediction or an aggregation scheme in a sensor network, concerning the number of transmissions and, eventually, the energy consumption levels. In this Section, we make a parameter study over the model parameters C , D , ρ , and α . Our goal is to observe how the number of transmissions varies in different scenarios and apply the model using simulations with normally distributed data.

5.1. Simulation setup

In OMNET++ [30], we simulated TelosB motes [31] using a TDMA-based Medium Access Control (MAC) protocol. In the MAC protocol adopted, each sensor node has a reserved slot to transmit. Therefore, we did not experience collisions during the transmissions, and there was no overhearing. We highlight that other MAC protocols may obtain different results, due to concurrent transmissions, although we can expect a similar reduction in their number of transmissions.

Regarding the mechanisms adopted to reduce the number of transmissions, we simulated three combinations: (i) with no prediction and no aggregation; (ii) with prediction, but no aggregation; and (iii) with aggregation, but no prediction. When data prediction was adopted, the prediction models were chosen in the GW, and *GW-to-node* transmissions were always aggregated.

As we showed before, in monitoring applications with DPSs, the number of transmissions is highly affected by the correlation between measurements made by the sensor nodes in a *sub-tree*, and by the predictions' accuracy. Therefore, regarding the model parameters, we observed the impact of different values of ρ , α . Values of ρ varied among 0.1, 0.2, ..., 0.9, and 0.95, and values α varied among 0.5, 0.7, 0.9, and 0.95.

Recall that, according to (37), the number of transmissions does not depend on the density of sensor nodes (C), but on the number of rings (D). Thus, to observe the impact of the growth in the number of sensor nodes in WSNs, we observed the number of transmissions with values of D varying among 1, 2, ..., and 10. Note that, when new rings are added, the number of sensor nodes increases quadratically if no aggregation is adopted. However, the number of transmissions does not change if sensor nodes aggregate them.

Finally, in our simulations, sensor nodes made one measurement per minute ($f = 1/60$), and the GW predicted their measurements once a day during three days ($T = 3 \times 86400$ seconds). Therefore, each sensor node made 4320 measurements, from which 1440 happened between each prediction model choice (in the cases when predictions were adopted).

Data: n = number of nodes, α = accuracy, ρ = correlation

Result: $P(n, \alpha, \rho)$ = probability that no transmission happens

```

1 if  $n = 0$  then
2   return  $P \leftarrow 1$ 
3 else
4    $q \leftarrow \lceil \Phi^{-1} \left( \frac{1-\alpha}{2} \right) \rceil$ 
5    $Q \leftarrow \{q, q, \dots, q\}_{1 \times n}$ 
6    $Y \leftarrow \{Y_1, Y_2, \dots, Y_n\}$ 
7    $\Sigma \leftarrow \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}_{n \times n}$ 
8   return  $P \leftarrow \Phi(Y, \Sigma, Q)$ 
9 end

```

Algorithm 1: Algorithm to calculate the probability that no transmission will be made.

5.2. Simulated algorithm

Assuming normally distributed values, the expected number of transmissions and receptions can be estimated using the cumulative density functions of MVN distributions. Based on (32), we designed the algorithm described in Algorithm 1. It calculates the probability P of making no transmissions in a group of n sensor nodes measuring data with correlation ρ if the average predictions' accuracy is α .

We highlight that, in our model, the number of children is used to define how many distributions will be used, which means that decimal values cannot be considered. Hence, we rounded all of them up to the next integer, which resulted on an upper bound for the number of transmissions in the simulations.

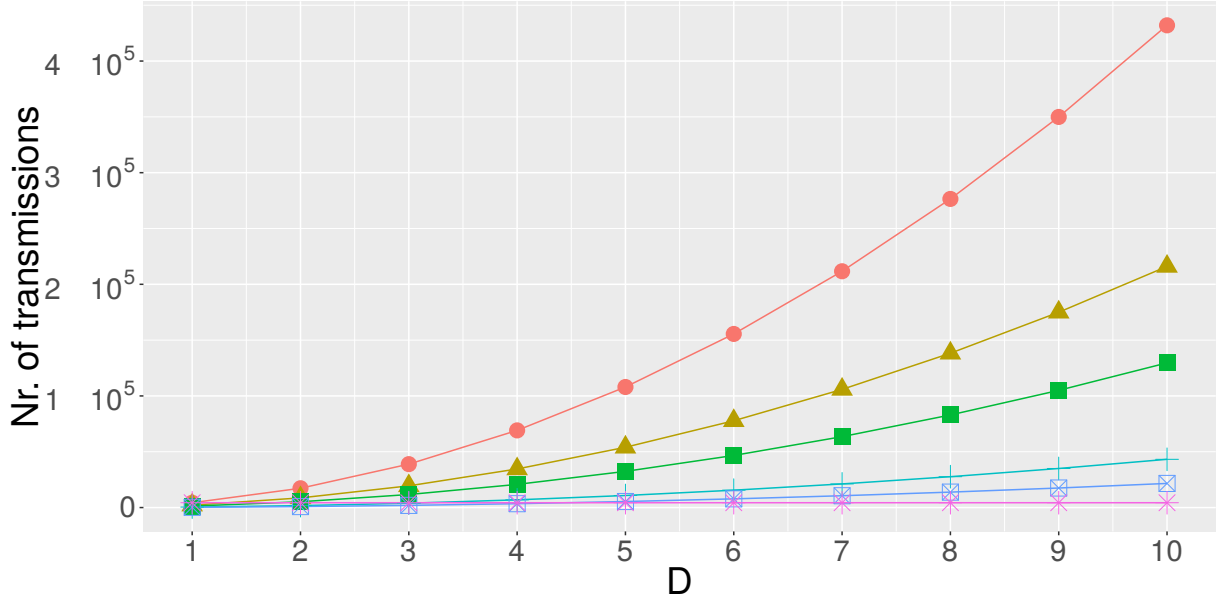
5.3. Number of transmissions

Given that the bottlenecks of a sensor network are the sensor nodes in the first ring, we calculate the number of transmissions at a sensor node i in ring $d = 1$ as

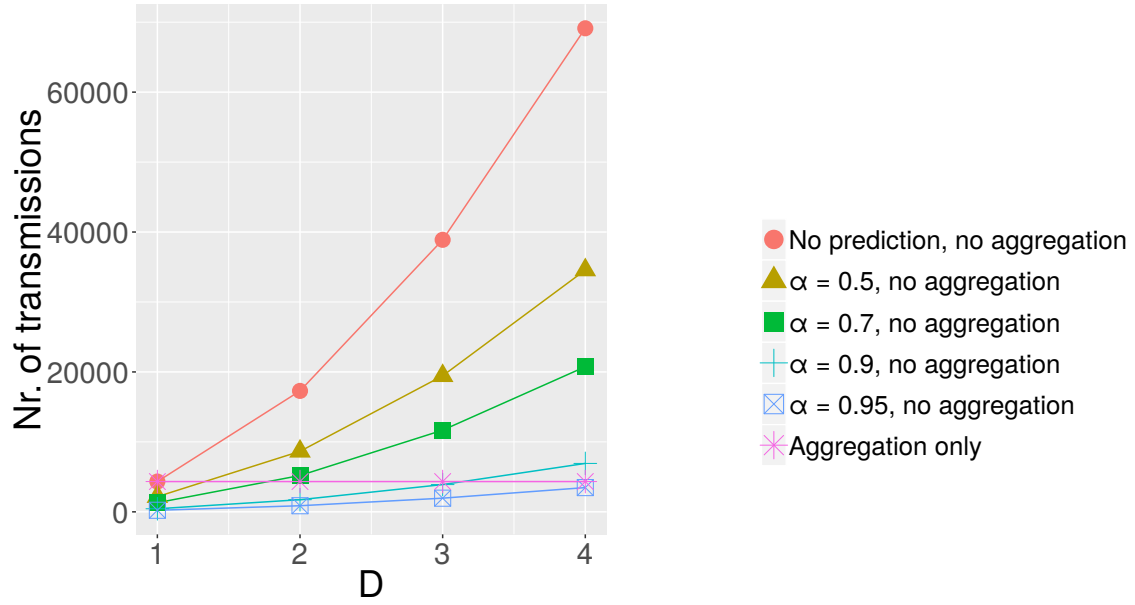
$$E(S''_{i,1}) = ((1 - P(K_1, \alpha, \rho)) f) + I_1) T, \quad (38)$$

and the number of receptions as

$$E(R''_{i,1}) = (1 - P(K_1, \alpha, \rho)) I_1 f T. \quad (39)$$



(a) The aggregation reduces the number of transmissions from quadratic to linear order.



(b) When number of rings is small ($D \leq 4$), the use of predictions can lead to fewer transmissions than the aggregation scheme.

Figure 7: The impact of the network size in the number of transmissions in the first ring.

Figure 7 shows the results for all the tested configurations. In larger sensor networks ($D > 4$), data aggregation has a higher impact than data prediction in the number of transmissions, as shown in Figure 7a. Similar results were observed in another study [10], but the authors did not realize that the predictions had less impact in the final savings and concluded that such optimal achievements happened due to the high accuracy of the predictions.

When the predictions are highly accurate, and the number of rings is small ($D \leq 4$), the data prediction has a higher impact on the number of transmissions if compared with the scenarios where the data is only aggregated. Figure 7b highlights scenarios with less than five rings.

To detail the power of the prediction and aggregation schemes, we considered a sensor network with five rings in which the aggregation scheme could reduce to 12.5% the number of transmissions, similarly to the most accurate predictions. Figure 8 highlights the gains obtained by adopting both schemes, where 100% of transmissions represent the case where only data aggregation is adopted (i.e., 12.5% of the transmissions in a WSN with no optimization). First, we can observe that the number of transmissions can be reduced to 15% of its maximum in the best scenario, where the predictions are highly accurate, and the measurements in the *sub-tree* are highly correlated. This represents around 1.88% of the total number of transmissions, given that it is 15% of the 12.5% of transmissions that the data aggregation would do.

Additionally, we did not observe any significant gains when the predictions were less accurate (around 0.5) nor when the predictions were more accurate (around 0.7), and the correlations were less than 0.7. Finally, with an average correlation (0.5), increasing the accuracy from 0.5 to 0.9 reduced by 30% the number of transmissions. Meanwhile, with an accuracy of 0.5, increasing the correlation from 0.5 to 0.9 reduced only by 6.5% the number of transmissions, which illustrates that the impact of making accurate predictions is much higher than having a high correlation between the measurements.

5.4. Energy consumption

Based on the number of transmissions and receptions, we can use the model presented before to estimate the total energy consumption of a sensor node i in ring d as

$$E(\text{En}_{i,d}'') = E(S_{i,d}'')\text{En}_{\text{TX}} + E(R_{i,d}'')\text{En}_{\text{RX}} + \text{En}_{\text{DIS}} + \text{En}_{\text{MIN}}, \quad (40)$$

where En_{TX} and En_{RX} are the extra energy consumption to respectively transmit and receive one packet, En_{MIN} is the minimum energy necessary to keep sensor nodes working without transmitting and receiving anything, and En_{DIS} depends on where the prediction models are chosen. If prediction models are independently chosen, $\text{En}_{\text{DIS}} = 0$; if prediction models are chosen in the GW, $\text{En}_{\text{DIS}} = \text{En}_{\text{RX}} + I_d \text{En}_{\text{TX}}$; and if prediction models are chosen in sensor nodes, $\text{En}_{\text{DIS}} = I_d \text{En}_{\text{RX}} + \text{En}_{\text{TX}}$.

To illustrate the applicability of this model, we estimated the energy consumption in a WSN after three days of operation and compared with the results obtained in our simulations. For this estimation, we

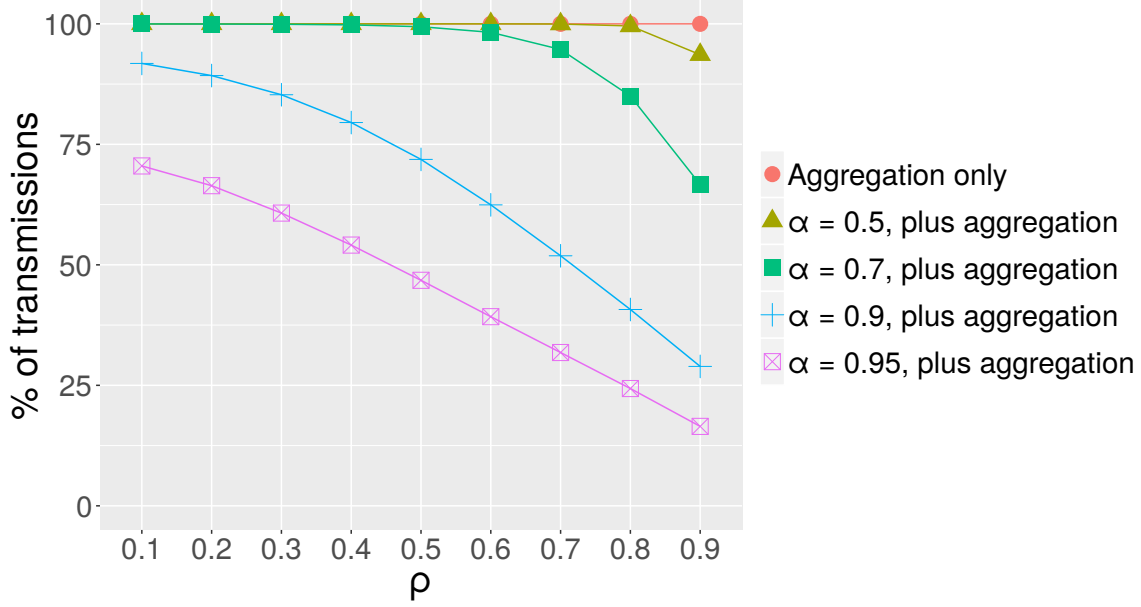


Figure 8: The effectiveness of the aggregations depend on the correlation between the measurements in a *sub-tree*.

considered a homogeneous sensor network with $D = 5$ and $C = 3$ (i.e., 75 sensor nodes plus the GW). To obtain the values of En_{TX} , En_{RX} , and En_{MIN} , we simulated three TelosB motes in OMNET++ transmitting and receiving data without making any predictions. After one simulated day, we calculated the average values for each parameter.

So far, we did neither distinguish delays nor packet lengths used in aggregated transmissions and receptions from the case without aggregation. In fact, in a real implementation, these transmissions could be done in the same packet types if we adopted simple aggregation functions, such as the maximum, minimum and the average of the measurements. However, larger packets would mean higher energy consumption to transmit and receive, in comparison with the non-aggregated transmissions. Therefore, to show the extensibility of our model, we used packets with eight times the payload of the normal packets in the aggregated transmissions.

To illustrate the results, we focused on the energy consumption of a sensor node in the first ring. As sensor nodes in the first ring must handle the highest number of transmissions, they consume more energy than the others. As a consequence of such an energy consumption, these sensor nodes can run out of battery earlier than those in the other rings, which has a substantial impact on the WSNs' lifetime. In Figure 9, we can see that just adopting the aggregation scheme (without making predictions) reduces the extra energy consumption to 60% of the total, yet larger packets are used. The greatest gains, nonetheless, are obtained after adopting the DPS and the aggregation scheme: they can save up to nearly 92% of the energy consumed by the transmissions. As explained before, the predictions' accuracy is more significant and has a higher

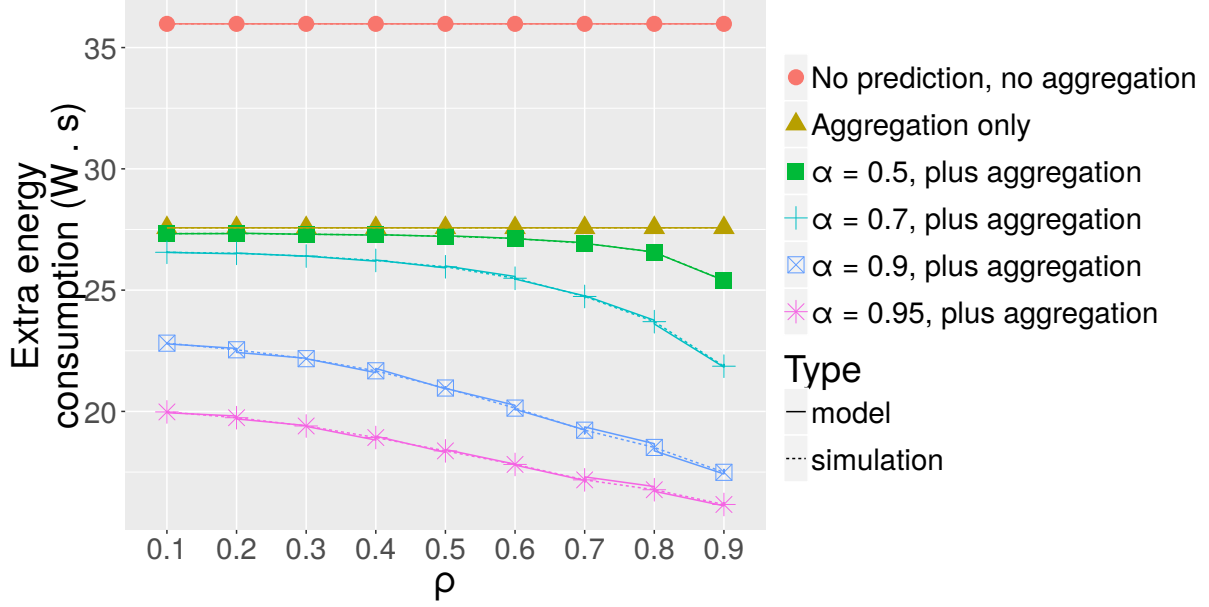


Figure 9: The model provides reliable results when compared with the simulations.

impact than the correlation between the measurements in a *sub-tree*. Hence, a very low correlation (0.1) with highly accurate predictions (0.95) give better results than a high correlation (0.9) with an average accuracy (0.5).

In fact, regardless of the values shown in the plot, the exact amount of saved energy depends on the hardware of the sensor nodes, their Operating System (OS), and the MAC protocol in use, besides other configurations. Nonetheless, the consumption is mainly driven by the relation between the minimum energy necessary to keep a sensor node making measurements and the amount of battery required for transmitting and receiving a packet. In conclusion, the results presented here can facilitate the decision about adopting a DPS in a WSN with a similar arrangement, even if the sensor nodes' configurations differ from those considered in our investigation.

6. Model validation

In Section 5, we designed an algorithm to calculate the impact of the data characteristics in the number of transmissions in scenarios where the WSN's structure follows the transmission model presented in Section 4. In this Section, however, our goal is to study a real use case, where the network is not as uniform as we assumed before, and the data is not exactly normally distributed but collected by real wireless sensor nodes in an experiment.

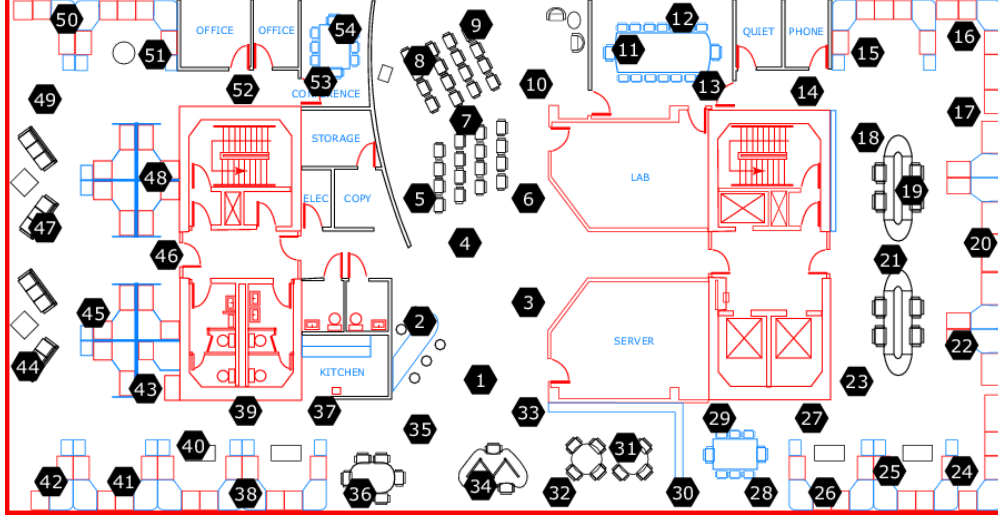


Figure 10: In the experiments made in the Intel Lab, temperature, relative humidity, and luminance were measured by 54 sensor nodes arranged as shown in this map [32].

6.1. Data selection

To validate the model, we adopted the data collected in the Intel Lab in the experiments described by Bodik et al. [32] and illustrated in Figure 10. In that experiment, 54 sensor nodes transmitted measurements of temperature, relative humidity, and luminance every 30 seconds during 37 consecutive days.

Because there were several missing values in the original dataset, we generated a new dataset by selecting a random measurement of each sensor in every interval of five minutes for eight days in a row. We randomly selected nine sensor nodes that reported in more than 95% of the time: 21, 22, 26, 31, 38, 40, 42, 45, and 46. In this new scenario, the network is not as uniform as we assumed before, which challenges the applicability of our model.

From the new dataset, we selected eight consecutive days to calculate the data characteristics, i.e., the average and the standard deviation of the temperature measurements made by each sensor in each hour of each day. To estimate the average correlation coefficient between the measurements made by these sensor nodes, we first calculated the correlation matrix computed for all possible pairs of sensor nodes, applied the Fisher’s z-transformation to transform these correlations into z-scores, calculated the average of the values and, finally, backtransformed the average to an r-score [33]. For the temperature values, we calculated an average correlation coefficient of 0.820068.

Based on the average and the standard deviation values, we used the z-table to calculate four acceptance thresholds based on the expected accuracy levels: 0.5, 0.7, 0.9, and 0.95. After calculating the average and four acceptance thresholds per sensor node per hour of each day, we considered that if a measurement differed from the average by less than the acceptance threshold, the node would not have transmitted during

those days.

We highlight that our focus at this point is to compare the results obtained with the model and not how to generate prediction models. In practice, it would be necessary a data study over the days preceding these experiments to design prediction models that would accurately predict the temperature measurements. Therefore, due to a lack of space, we assume that those prediction models could have been computed using state-of-the-art algorithms for predictions and the data collected in the period before the experiments. Given that, the GW could disseminate these models, for example, some hours before the day that the predictions would refer.

Finally, we calculated the total number of transmissions made by a sensor node that would be connected to the GW and compared it with the case in which the sensor nodes would transmit every five minutes regardless the predictions. Assuming a deployment where only one of the selected sensor nodes was directly connected to a GW, every transmission would go through that sensor node, and there would be necessary 20736 transmissions to send all measurements to the GW during the eight days considered in the experiment.

6.2. Results

Recall the WSN model shown in Section 4, where each branch is connected to the GW via one sensor node. It was possible to infer the number of transmissions at that sensor node considering only the number of children it would have. Such a value was defined as K_1 and depends on the number of rings in the network. For example, in a WSN with D rings, $K_1 = D^2 - 1$. Therefore, given that we are working with $K_1 = 8$ sensor nodes, a WSN with three rings would have the same number of transmissions between the GW and the sensor node which is closest to the GW. In other words, the nine sensor nodes selected for our observations could represent a branch in the model presented before with $D = 3$ and $C \geq 3$. Given that every branch in the model is expected to have the same number of sensor nodes and transmissions, the percentage of saved transmissions will be also similar to the percentage of saved transmissions calculated for a complete WSN.

In an optimized deployment where a sensor node aggregates measurements from its children and transmits only once every five minutes, it would be necessary 2304 transmissions to send all measurements to the GW during the eight days considered in the experiment.

Alternatively, if we adopted the DPS besides the aggregation scheme, each transmission would be conditioned to the inaccuracy of (at least) one prediction in that transmission's five-minute window. Table 2 shows the number of temperature measurements transmitted using DPS with aggregation, considering a deployment with aggregation as a reference (100%) for the maximum number of transmissions.

As expected, we could observe some difference between the model and the number of transmissions expected in the real scenario. We believe that this variation is mainly because the model is built based on the assumption the data is normally distributed, which was not true in the real experiment. This

Accuracy	Real data	Model	Difference
50%	87.3%	92%	4.6%
60%	78.8%	81.1%	2.3%
70%	67.4%	66.4%	-1.1%
80%	51.1%	48.7%	-2.4%
90%	27.6%	28%	0.5%
95%	11.2%	15.8%	4.5%

Table 2: Total number of transmissions considering different predictions’ accuracy levels for the temperature measurements.

change impacts the final number of transmissions and partly explains the differences between the number of transmissions obtained using the data from the real experiment and the model. Apart of that, the model seems still valid as a close approximation for this real case because the absolute difference between the results obtained using real data and those calculated using the model were smaller than 2.5% in most of the cases and smaller than 5% in the worst case.

7. Related work

The mechanism called BBQ assumes that sensors nearby are likely to have correlated readings, which may mean that most of the measurements provide little benefit in approximate answers’ quality [26]. BBQ approximates the probability density function of the measurements to MVN distributions and, given the correlation between known measurement(s) and the unknown one(s), it calculates their expected value associated with a confidence interval. If the confidence level is greater or equal than a user defined threshold, it answers the queries locally, without triggering new transmissions from sensor nodes. In an outdoor scenario with less interference from humans and machines (where the sensed data approximates better to Normal distributions), it was possible to achieve a significant reduction in the number of transmissions. That improvement represented a reduction of 97% in the energy consumption of the nodes and an acceptable level of mistakes (nearly 5% of wrong answers).

Similarly to the model that we propose, the authors incorporated statistical models of real-world processes and exploited the correlation between measurements taken in the same vicinity. They also incorporated the correlation between different types of data that the sensor nodes may be able to measure, for example, their voltage and the local temperature. On the other hand, the authors focused on modeling the number of transmissions in existing WSNs to optimize their mechanism of query answering. Our model is intended to model generic WSNs used to report measurements periodically, based on their expected characteristics and takes into account the relative importance of nodes close to the GW, i.e., those that must handle the highest

number of transmissions and, therefore, are the bottleneck of WSNs' workload.

Intanagonwivat et al. have already modeled the impact of data aggregation schemes in WSNs [34]. More specifically, they showed the impact of network density on data aggregation using directed diffusion with a greedy algorithm to construct trees. In their work, they proposed the in-network aggregation (along with the greedy tree) and compared the perfect aggregation (which saves the headers and merges the content in new packets with the same length) with the linear aggregation (which also saves the headers, but appends the content in larger payloads). According to simulation results, the greedy algorithm can reduce 36% of the energy consumption using linear aggregation and 43% using the perfect aggregation in high-density networks. Finally, the authors concluded that, in high-density networks, more energy is saved by the greedy algorithm, and the delays are as good as using the opportunistic algorithm.

More recently, Fan et al. focused on calculating the most energy efficient deployment strategy for WSNs using an integral programming model [35]. They used a regular hexagonal cell architecture [36, 37] to model the location of sensor nodes in the plane, which, in fact, is similar to a ring model with $C = 6$. Using this model, they formulated the energy consumption of sensor nodes and GWs based on the energy used to transmit, receive and process data in sensor nodes. Finally, they also evaluated the impact of data aggregation in the energy consumption of the sensor nodes.

The works above inspired our model for data transmissions. However, different from our proposal, they focused only on data aggregation and did not evaluate the effects of DPSs in WSNs. Concerning the evaluation of DPSs, Liu et al. were the first authors to introduce statistical methods to choose which prediction model better fits a certain environment [22]. They created a formula to estimate the Prediction Cost (PC), which considers the percentage of transmitted measurements (r) and the user desired level of accuracy (α) [22]. More recently, an extended formula was designed and implemented in real sensor nodes to compare the savings using several prediction methods, such as the *Constant*, ES and ARIMA methods [38, 39]. The new formula is more generic than the original proposed and also considers the prediction models' memory footprint (Ec) as a significant computational cost for sensor nodes:

$$PC = [\alpha f(e) + (1 - \alpha)r] Ec, \quad (41)$$

where e is the measure of the predictions' accuracy (e.g., Mean Square Error (MSE), Root Mean Square Error (RMSE), symmetric Mean Absolute Percentage Error (sMAPE)) and $f(e)$ is the accuracy according to the chosen measure. The formula of PC can be useful to decide for adopting a DPS or not, but it is limited to evaluate if predictions will save energy in one sensor node and does neither consider the impact of occasional aggregations, nor the work to forward transmissions from children nodes.

To the best of our knowledge, this is the first work that models the impact of DPSs in sensor networks and shows their potential as a whole class of applications.

8. Conclusion and future work

In this work, we presented a mathematical framework to calculate the gains and benefits of adopting a DPS to reduce the number of transmissions in a WSN. The model shows that, concerning the number of transmissions, the benefits of adopting an aggregation scheme are greater than using only predictions, and that combining both leads to the highest savings. For example, as observed in Figure 8, the number of transmissions and receptions in the bottlenecks (i.e., the sensor nodes in the first ring) can be reduced by nearly 98% using accurate predictions and data aggregation. Our simulations also showed that the accuracy of the predictions impacts more than the correlation between the measurements from different sensor nodes. Finally, we compared the values obtained using our model with those that would have been obtained using real data, and we observed that our model provides a good estimation of the number of transmissions that might happen in practice.

The main contribution for the future generations of WSNs is a model that relies on the statistical theory to show the impact of sensor nodes' hardware evolution and the integration of WSNs into the IoT, which can be exploited to make predictions with higher accuracy in DPSs. The backbone of the model consists of an application of two statistical theorems: (i) the Central Limit Theorem, which supports the normalization of the data measured by the sensor nodes; and (ii) the Law of Large Numbers, which allows the extrapolation to similar scenarios based on the average number of neighbors and furthest distance from sensor nodes to GWs.

Considering that this model has been designed to represent different types of sensor networks, there are some challenges to set up the best parameters for each use case. From our experiments, we expect that the critical points in a real scenario may vary between: (i) finding the precise correlation between measurements from several sensor nodes; (ii) approximating the measurements to Normal distributions, since it may require some data analysis in advance; (iii) having restrictions about changing the radio operation of some sensor nodes, because they may be fundamental to the sensor network connectivity, and the timeliness of the data delivered by the WSN; and (iv) calculating the energy necessary for each step (i.e., transmission, reception, etc.), because details in the software and hardware implementations may influence such values [40].

In future works, we plan to include this model in a larger system that can improve the accuracy of the predictions made by the GW after consulting external sources of information. With such an extended access to external information and using the existing formulas to estimate the prediction costs [22, 38, 39], GWs will be able to choose the best mechanism to evaluate the quality of the measurements provided by their WSNs. Hence, all these features will make it possible to run self-managed systems that adapt the sensor network's operation according to their surroundings (and not only based on the observed environment) to achieve the best results, i.e., highest quality of measurements and fewest transmissions possible.

Acknowledgment

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502), TEC2012-32354 (MINECO/FEDER) and by the ENTOMATIC FP7-SME-2013 EC project (605073).

Bibliography

References

- [1] R. Jurdak, A. Elfes, B. Kusy, A. Tews, W. Hu, E. Hernandez, N. Kottege, P. Sikka, Autonomous surveillance for biosecurity, *Trends in Biotechnology* 33 (4) (2015) 201–207. arXiv:0210091v1, doi: 10.1016/j.tibtech.2015.01.003.
URL <http://dx.doi.org/10.1016/j.tibtech.2015.01.003>
- [2] D. J. McCorrie, E. Gaura, K. Burnham, N. Poole, R. Hazelden, *Predictive Data Reduction in Wireless Sensor Networks Using Selective Filtering for Engine Monitoring*, Springer New York, New York, NY, 2015, pp. 129–148. doi:10.1007/978-1-4939-2468-4_6.
URL http://dx.doi.org/10.1007/978-1-4939-2468-4_6
- [3] J. A. Stankovic, Research challenges for wireless sensor networks, *ACM SIGBED Review* 1 (2) (2004) 9–12. doi:10.1145/1121776.1121780.
URL <http://dl.acm.org/citation.cfm?id=1121780>
- [4] L. M. Borges, F. J. Velez, A. S. Lebres, Survey on the Characterization and Classification of Wireless Sensor Network Applications, *IEEE Communications Surveys & Tutorials* 16 (4) (2014) 1860–1890. doi:10.1109/COMST.2014.2320073.
URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6805127>
- [5] K. Zheng, H. Li, The Transmission Strategy for Energy Harvesting Wireless Transmitters, in: 2011 IEEE Global Telecommunications Conference - GLOBECOM 2011, Vol. 2, IEEE, 2011, pp. 1–5. doi: 10.1109/GLOCOM.2011.6134169.
URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6134169>
- [6] J. P. Amaro, F. J. T. E. Ferreira, R. Cortesao, J. Landeck, Energy harvesting for Zigbee compliant Wireless Sensor Network nodes, in: *IECON 2012 - 38th Annual Conference on IEEE Industrial Electronics Society*, IEEE, 2012, pp. 2583–2588. doi:10.1109/IECON.2012.6388844.
URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6388844>

- [7] T. Adame, A. Bel, B. Bellalta, J. Barcelo, M. Oliver, IEEE 802.11AH: the WiFi approach for M2M communications, *IEEE Wireless Communications* 21 (6) (2014) 144–152. doi:10.1109/MWC.2014.7000982.
URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7000982>
- [8] Y. Chang, C. Zhou, Coordinated Random Access Management for Network Overload Avoidance in Cellular Machine-to-Machine Communications, in: *European Wireless 2014; 20th European Wireless Conference; Proceedings of, 2014*, pp. 627–632.
- [9] K. Chatzikokolakis, A. Kaloxylos, P. Spapis, N. Alonistioti, C. Zhou, J. Eichinger, . Bulakci, On the way to massive access in 5g: Challenges and solutions for massive machine communications, in: M. Weichold, M. Hamdi, M. Z. Shakir, M. Abdallah, G. K. Karagiannidis, M. Ismail (Eds.), *Cognitive Radio Oriented Wireless Networks*, Vol. 156 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, Springer International Publishing, 2015, pp. 708–717. doi:10.1007/978-3-319-24540-9_58.
URL http://dx.doi.org/10.1007/978-3-319-24540-9_58
- [10] S. Santini, K. Römer, An Adaptive Strategy for Quality-Based Data Reduction in Wireless Sensor Networks, in: *3rd International Conference on Networked Sensing Systems*, 2006, pp. 29 – 36.
URL http://vs.inf.ethz.ch/publ/papers/santinis_inss2006.pdf
- [11] I. Lazaridis, S. Mehrotra, Capturing sensor-generated time series with quality guarantees, in: *Data Engineering, 2003. Proceedings. 19th International Conference on*, 2003, pp. 429–440. doi:10.1109/ICDE.2003.1260811.
URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1260811
- [12] B. R. Stojkoska, D. Solev, D. Davcev, Data prediction in WSN using variable step size LMS algorithm, in: *Proceedings of the 5th International Conference on Sensor Technologies and Applications*, 2011.
- [13] M. Wu, L. Tan, N. Xiong, Data prediction, compression, and recovery in clustered wireless sensor networks for environmental monitoring applications, *Information Sciences* 329 (2016) 800–818. doi:10.1016/j.ins.2015.10.004.
URL <http://dx.doi.org/10.1016/j.ins.2015.10.004>
- [14] Y.-A. Le Borgne, S. Santini, G. Bontempi, Adaptive model selection for time series prediction in wireless sensor networks, *Signal Process.* 87 (12) (2007) 3010–3020. doi:10.1016/j.sigpro.2007.05.015.
URL <http://dx.doi.org/10.1016/j.sigpro.2007.05.015>
- [15] G. Li, Y. Wang, Automatic ARIMA modeling-based data aggregation scheme in wireless sensor networks, *EURASIP Journal on Wireless Communications and Networking* 2013 (1) (2013) 85. doi:

10.1186/1687-1499-2013-85.

URL <http://jwcn.eurasipjournals.com/content/2013/1/85>

- [16] D. J. McCorrie, E. Gaura, K. Burnham, N. Poole, R. Hazelden, Predictive data reduction in wireless sensor networks using selective filtering for engine monitoring, in: *Wireless Sensor and Mobile Ad-Hoc Networks: Vehicular and Space Applications*, Springer New York, New York, NY, 2015, pp. 129–148. doi:10.1007/978-1-4939-2468-4_6.
URL http://dx.doi.org/10.1007/978-1-4939-2468-4_6
- [17] R. Askari Moghadam, M. Keshmirpour, Hybrid arima and neural network model for measurement estimation in energy-efficient wireless sensor networks, in: A. Abd Manaf, S. Sahibuddin, R. Ahmad, S. Mohd Daud, E. El-Qawasmeh (Eds.), *Informatics Engineering and Information Science*, Vol. 253 of *Communications in Computer and Information Science*, Springer Berlin Heidelberg, 2011, pp. 35–48. doi:10.1007/978-3-642-25462-8_4.
URL http://dx.doi.org/10.1007/978-3-642-25462-8_4
- [18] J. Kho, A. Rogers, N. R. Jennings, Decentralized control of adaptive sampling in wireless sensor networks, *ACM Transactions on Sensor Networks* 5 (3) (2009) 1–35. doi:10.1145/1525856.1525857.
URL <http://portal.acm.org/citation.cfm?doid=1525856.1525857>
- [19] Y. Chen, Z. Wang, Z. Zhao, D. Li, L. Cui, Online model-driven data acquisition for wireless sensor networks, in: *2015 IEEE Wireless Communications and Networking Conference (WCNC)*, IEEE, 2015, pp. 1572–1577. doi:10.1109/WCNC.2015.7127702.
URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7127702>
- [20] S. Cheng, Z. Cai, J. Li, X. Fang, Drawing dominant dataset from big sensory data in wireless sensor networks, in: *2015 IEEE Conference on Computer Communications (INFOCOM)*, no. 1, IEEE, 2015, pp. 531–539. doi:10.1109/INFOCOM.2015.7218420.
URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7218420>
- [21] S. Goel, T. Imielinski, Prediction-based monitoring in sensor networks: taking lessons from MPEG, *ACM SIGCOMM Computer Communication Review* 1.
URL <http://dl.acm.org/citation.cfm?id=1037117>
- [22] C. Liu, K. Wu, M. Tsao, Energy efficient information collection with the ARIMA model in wireless sensor networks, in: *IEEE Global Communications Conference*, 2005, pp. 2470–2474.
URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1578206

- [23] G. M. Dias, T. Adame, B. Bellalta, S. Oechsner, A self-managed architecture for sensor networks based on real time data analysis, in: 2016 IEEE Future Technologies Conference, IEEE, 2016.
URL <https://arxiv.org/abs/1605.09011>
- [24] H. Jiang, S. Jin, C. Wang, Prediction or Not? An Energy-Efficient Framework for Clustering-based Data Collection in Wireless Sensor Networks, IEEE Transactions on Communications 39 (12) (1991) 1721–1725. doi:10.1109/26.120156.
URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5601711
- [25] K. Langendoen, A. Meier, Analyzing mac protocols for low data-rate applications, ACM Trans. Sen. Netw. 7 (2) (2010) 19:1–19:40. doi:10.1145/1824766.1824775.
URL <http://doi.acm.org/10.1145/1824766.1824775>
- [26] A. Deshpande, C. Guestrin, S. R. Madden, J. M. Hellerstein, W. Hong, Model-driven data acquisition in sensor networks, in: Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30, VLDB '04, VLDB Endowment, 2004, pp. 588–599.
URL <http://dl.acm.org/citation.cfm?id=1316689.1316741>
- [27] J. S. Armstrong, Principles of forecasting: a handbook for researchers and practitioners, Springer Science & Business Media, 2001.
URL <https://www.gwern.net/docs/predictions/2001-principlesforecasting.pdf>
- [28] F. Wang, J. Liu, Networked Wireless Sensor Data Collection: Issues, Challenges, and Approaches, IEEE Communications Surveys & Tutorials 13 (4) (2011) 673–687. doi:10.1109/SURV.2011.060710.00066.
URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5497857>
- [29] A. Genz, Comparison of Methods for the Computation of Multivariate Normal Probabilities, Computing Sciences and Statistics 25 (1993) 400 – 405.
URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.9631&rep=rep1&type=pdf>
- [30] A. Varga, The OMNeT++ discrete event simulation system, Proceedings of the European Simulation Multiconference (ESM'2001) (S 185) (2001) 65.
URL <https://labo4g.enstb.fr/twiki/pub/Simulator/SimulatorReferences/esm2001-meth48.pdf>
- [31] Crossbow Technology Inc., TelosB - TelosB Mote Platform, revision B (2004). doi:6020-0094-01Rev.B.
URL http://www.willow.co.uk/TelosB_Datasheet.pdf

- [32] P. Bodik, W. Hong, C. Guestrin, S. Madden, M. Paskin, R. Thibaux, Intel lab data, online dataset (Jun. 2004).
URL <http://db.lcs.mit.edu/labdata/labdata.html>
- [33] N. C. Silver, W. P. Dunlap, Averaging Correlation Coefficients: Should Fisher's z Transformation Be Used?, *Journal of Applied Psychology* 72 (1) (1987) 146–148.
- [34] C. Intanagonwiwat, D. Estrin, R. Govindan, J. Heidemann, Impact of network density on data aggregation in wireless sensor networks, *Proceedings 22nd International Conference on Distributed Computing Systems* (2001) 457–458 doi:10.1109/ICDCS.2002.1022289.
URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1022289>
- [35] T. Fan, G. Teng, L. Huo, Deployment strategy of WSN based on minimizing cost per unit area, *Computer Communications* 38 (2014) 26–35. doi:10.1016/j.comcom.2013.10.002.
URL <http://dx.doi.org/10.1016/j.comcom.2013.10.002>
- [36] B. Li, Q. Wang, Y. Yang, J. Wang, Optimal Distribution of Redundant Sensor Nodes for Wireless Sensor Networks, in: *2006 IEEE International Conference on Industrial Informatics*, Vol. 00, IEEE, 2006, pp. 985–989. doi:10.1109/INDIN.2006.275731.
URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4053523>
- [37] H. Subir, G. Amrita, S. Sanjib, D. Avishek, D. Sipra, A lifetime enhancing node deployment strategy in WSN, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 5899 LNCS (2009) 295–307. doi:10.1007/978-3-642-10509-8_33.
- [38] F. A. Aderohunmu, G. Paci, D. Brunelli, J. D. Deng, L. Benini, Prolonging the lifetime of wireless sensor networks using light-weight forecasting algorithms, in: *2013 IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, IEEE, 2013, pp. 461–466. doi:10.1109/ISSNIP.2013.6529834.
URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6529834
- [39] F. A. Aderohunmu, G. Paci, D. Brunelli, J. D. Deng, L. Benini, M. Purvis, An Application-Specific Forecasting Algorithm for Extending WSN Lifetime, *2013 IEEE International Conference on Distributed Computing in Sensor Systems* (2013) 374–381 doi:10.1109/DCOSS.2013.51.
URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6569459
- [40] C. Pham, Communication performances of IEEE 802.15.4 wireless sensor nodes for data-intensive applications: A comparison of WaspMote, Arduino MEGA, TelosB, MicaZ and iMote2 for image surveillance,

Appendix A. Minimum average accuracy

Let us assume a monitoring WSN with homogeneous sensor nodes periodically transmitting measurements every $1/f$ seconds, and a sensor node i in ring d . According to (6), the number of transmissions at i in a period of $1/f$ seconds is the sum $E(S_i) + E(R_i)$. If a DPS is adopted, the number of transmissions will be $E(S'_i) + E(R'_i) + E(X_{\text{DIS}})$, as defined in (21). Therefore, a DPS will reduce the number of transmissions during a period T if and only if:

$$(E(S'_i) + E(R'_i))fT + E(X_{\text{DIS}}) \leq (E(S_i) + E(R_i))fT \quad (\text{A.1})$$

Thus, we can define α_{\min} as the minimum average between the accuracies of the children of i that would reduce the number of transmissions in a DPS (without aggregation). It must satisfy the following equation:

$$(E(S'_i) + E(R'_i))fT + E(X_{\text{DIS}}) = (E(S_i) + E(R_i))fT \quad (\text{A.2})$$

Based on (4), (5), (22), and (23), we can calculate it as:

$$((K_d + 1) \alpha_{\min}^c + (K_d \alpha_{\min}^c))fT + E(X_{\text{DIS}}) = ((K_{i,d} + 1) + K_d)fT \quad (\text{A.3})$$

Knowing this, we can work out the equation:

$$\begin{aligned} ((K_d + 1) \alpha_{\min}^c + (K_d \alpha_{\min}^c))fT &= ((K_d + 1) + K_d)fT - E(X_{\text{DIS}}) \\ \alpha_{\min}^c(D^2 + (D^2 - 1))fT &= (D^2 + (D^2 - 1))fT - E(X_{\text{DIS}}) \\ \alpha_{\min}^c(D^2 + (D^2 - 1))fT &= (D^2 + (D^2 - 1))fT - E(X_{\text{DIS}}) \\ \alpha_{\min}^c &= \frac{(D^2 + (D^2 - 1))fT - E(X_{\text{DIS}})}{(D^2 + (D^2 - 1))fT} \\ \alpha_{\min}^c &= 1 - \frac{E(X_{\text{DIS}})}{(D^2 + (D^2 - 1))fT} \\ \alpha_{\min} &= \frac{E(X_{\text{DIS}})}{(D^2 + (D^2 - 1))fT} \\ \alpha_{\min} &= \frac{E(X_{\text{DIS}})}{(2D^2 - 1)fT} \end{aligned} \quad (\text{A.4})$$

Given that we assume no aggregation, the value of $E(X_{\text{DIS}})$ is defined by (10):

$$E(X_{\text{DIS}}) = E(S^*) + E(R^*) = (2D^2 - 1). \quad (\text{A.5})$$

Therefore,

$$\alpha_{\min} = 1/fT \quad (\text{A.6})$$

Appendix B. Data model

A Normal distribution is characterized by its probability density function whose pattern is often encountered in several types of observations. According to the Central Limit Theorem, the sampling distribution of the mean of any independent random variable tends to be Normal, even if the distribution from which the average is computed is decidedly non-Normal. For example, it has been shown that environmental readings—such as temperature, light, and humidity—done by outdoor WSNs can be approximated to normal distributions if properly managed [26].

We will assume that a sensor network is composed of a set of sensor nodes S and each sensor node $i \in S$ is responsible for measuring a certain parameter from the environment, such that the set of observations follows a Normal distribution with mean μ_i and variance σ_i^2 . By convention, this is represented as $Y_i = N(\mu_i, \sigma_i^2)$. A prediction \bar{y}_i (for example, $\bar{y}_i = \mu_i$) can be calculated by the sensor node i and the GW. We define the acceptance threshold ε_i , i.e., the prediction is told to be correct if the real observation (y_i) is in the interval $[\bar{y}_i - \varepsilon_i, \bar{y}_i + \varepsilon_i]$.

Assuming that the data is normally distributed, the chances of observing a new value inside the accepted interval can be calculated by normalizing the value of ε_i , i.e., rewriting it in terms of the variance σ_i^2 . The normalized value of ε_i is represented by z_i as

$$z_i = \frac{\varepsilon_i - \bar{y}_i}{\sigma_i} \quad (\text{B.1})$$

Thus, in this case, the accuracy of the predictions (α_i) can be calculated based on the cumulative distribution function of the normal distribution:

$$\Phi_{\mu,\sigma}(x) = \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{x - \mu}{\sigma\sqrt{2}} \right) \right]. \quad (\text{B.2})$$

Again, according to the Central Limit Theorem, we assume unbiased predictions and normally distributed errors. Therefore, the percentage of observations that will fall outside the accepted interval is represented by the two-tailed Z-test (i.e., $2\Phi(-|z_i|)$), and α_i is

$$\alpha_i = 1 - 2\Phi(-|z_i|). \quad (\text{B.3})$$

By substituting the Equation B.1 into Equation B.3, we can observe that

$$\alpha_i = 1 - 2\Phi \left(- \left| \frac{\varepsilon_i - \bar{y}_i}{\sigma_i} \right| \right), \quad (\text{B.4})$$

which shows that the accuracy of the predictions depends on the acceptance threshold, the mean and the variance of the data.

Appendix C. Proof of $1 - \alpha^x \leq x(1 - \alpha)$

Let us assume two values α and x such that $\alpha \in [0, 1]$ and $x \geq 1$. We want to show that $1 - \alpha^x \leq x(1 - \alpha)$:

$$\begin{aligned} 1 - \alpha^x &\leq x(1 - \alpha) \\ 1 - \alpha^x &\leq x - \alpha x \\ -\alpha^x + 1 &\leq x - \alpha x \\ \alpha^x &\geq 1 - x + \alpha x \\ \alpha^x &\geq 1 + x(\alpha - 1) \end{aligned} \tag{C.1}$$

When $\alpha = 0$ or $\alpha = 1$, we can easily observe that the affirmation is true because of $x \geq 1$ by definition. For the other values of α , we can use the Bernoulli's inequality:

$$(1 + i)^j \geq 1 + ij, \tag{C.2}$$

where $i > -1$, $i \neq 0$ is a real number and $j \geq 2$ an integer value. Substituting the values of α and x in Equation C.1 respectively by $i + 1$ and j , the claim is proved.