Smart Application-aware IoT Data Collection

Vasilios A. Siris, Nikos Fotiou, Alexandros Mertzianis, George C. Polyzos

Mobile Multimedia Laboratory, Department of Informatics

School of Information Sciences & Technology

Athens University of Economics and Business, Greece

{vsiris, fotiou, mertzianis15, polyzos}@aueb.gr

Abstract

We present and experimentally evaluate procedures for efficient IoT data collection while achieving target requirements in terms of data accuracy, response time, energy, and privacy protection. Different strategies are considered because different IoT applications can have different requirements. Specifically, the accuracy-driven strategy adjusts the time period between consecutive measurements following an additive increase and multiplicative decrease (AIMD) scheme based on a target data accuracy, while the time-driven strategy adjusts the time period between measurement requests to achieve delay less than a given maximum delay between consecutive measurements. The energy-driven strategy considers both the data accuracy and the energy costs for the corresponding measurements. Finally, the privacy-driven strategy adds noise to measurements using differential privacy techniques. The experimental evaluation involves real temperature, humidity, and ozone (O3) measurements obtained from three testbeds through the FIESTA-IoT platform. Our results show that the AIMD adaptation of the measurement period is robust to different types of measurements from different testbeds, without having any tuning parameters. Also, the experimental results show the tradeoffs between the target data accuracy and the number of measurements and between the target data accuracy and the corresponding energy costs. For the privacy-driven strategy, the results show that the addition of noise to the sensor measurements using differential privacy has a negligible effect on the aggregate statistics.

Keywords: data accuracy, adaptive data collection, differential privacy, testbed experiments

I. Introduction

The Internet of Things (IoT) will involve a huge number of sensors. Periodically collecting data from all IoT sensors wastes a significant amount of communication and storage resources, in addition to a significant amount of energy, which impedes the scalability of IoT systems.

Moreover, the collected data, or even correlating the collected measurements, may reveal sensitive information related to end-user activities. Naive approaches to (big) IoT data collection - reflecting the “first collect everything then (try to) analyse everything” paradigm - can also impede the integration of the IoT and Cloud systems and of the IoT and big data analysis, since the immense amount of data has implications on the amount of network resources and the amount of computation resources that are required.

Different IoT applications have different requirements in terms of accuracy, latency, and energy consumption. For example, an environmental monitoring application can require monitoring of real world phenomena with some degree of accuracy, while being tolerant to delays in receiving data updates from the IoT sensors. On the other hand, time-critical applications, such as security and critical infrastructure management that involve both monitoring and actuation, can have strict requirements in terms of the delay for IoT sensor nodes to transmit their data to the applications. Finally, different IoT sensors can have different constraints in terms of battery consumption hence applications can require a different balance between data accuracy or timeliness and energy consumption. Motivated by the above, the goal of our work in this paper is to develop and experiment with procedures for efficiently collecting IoT data while achieving target application requirements in terms of data accuracy, timeliness, energy efficiency, and privacy protection.

One approach to reduce the amount of IoT data collected is to perform filtering or aggregation at the sensors or at the edge of the IoT network close to the sensors. However, this might not always be possible since sensors typically have limited processing and storage resources, and implementing application policies at edge devices (e.g. gateways) close to the sensors is not always possible. This is the case with federated IoT systems, such as the platform developed by the EU-funded Horizon 2020 FIESTA-IoT project[1], where the sensors and their access networks have a different administration than the measurement platform with which applications interact to request and receive measurements. For the above reasons, in this paper we focus on data collection procedures that are implemented solely at the receiver-side, i.e. at the applications requesting IoT data.

In summary, the contributions of the paper are the following:

- We define procedures for efficient data collection that satisfy target requirements in terms of data accuracy, time, energy, and privacy protection. The data accuracy-driven procedure

[1] FIESTA-IoT: Federated Interoperable Semantic IoT Testbeds and Applications, http://fiesta-iot.eu/

adapts the period between measurements using an additive increase and multiplicative decrease (AIMD) scheme and has no tuning parameters. The time-driven strategy tries to ensure the timeliness of the last measurement received. The energy-driven strategy jointly considers the data accuracy and the energy costs. Finally, the privacy-driven procedure is based on differential privacy techniques.

- We evaluate the procedures with experiments involving real temperature, humidity, and ozone (O3) measurements, that are obtained from three testbeds over the FIESTA-IoT platform, which provides uniform access to IoT data from heterogeneous testbeds.

An important feature of the proposed AIMD scheme for adjusting the interval between consecutive measurements in the data accuracy-driven policy is that it does not have tuning parameters, as other model-based approaches in the literature, which we review in Section IV. Despite not having tuning parameters, our experimental evaluation shows that the AIMD scheme is very robust to measurements for different phenomena and from different testbeds.

The remainder of the paper is structured as follows: In Section II we present the overall architecture of our measurement framework, and describe the four data collections strategies that it supports. In Section III we present and discuss our experimental results. In Section IV we present a summary of related work, identifying how the work in this paper differs. Finally, in Section V we conclude the paper indicating directions of ongoing and future research.

II. IoT data collection strategies

Our data collection framework, which we will refer to as BeSmart, implements the following four strategies:

- Data accuracy-driven: This strategy considers the tradeoff between the data accuracy and the frequency of measurement requests. Specifically, the frequency of measurement requests is adapted (temporal adaptation) while maintaining a target data accuracy. This strategy can be extended when there are many IoT sensors located in the same geographic area: the collector can exploit the spatial correlation of sensor measurements to adapt the subset of the sensors from which measurements are requested (spatial adaptation).
- Time-driven: This strategy tries to ensure that the elapsed time since the timestamp of the last measurement is below some maximum delay; this elapsed time corresponds to the timeliness of data measurements. For the data collector to select the appropriate time to request measurements in order to ensure the maximum elapsed time target, the delay from

the time a request is sent by the collector until the time the response (measurement) is received must be considered; the latter delay can involve the network delay, the delay of the IoT measurement middleware, and the delay at the particular IoT testbed where the sensor providing the data measurements is located.

- Energy-driven: This strategy seeks to maximize the net benefit, which is the gain for a given data accuracy or timeliness minus the corresponding power consumption. The gain for a specific data accuracy or timeliness can be expressed using a utility function, while the power consumption depends on the frequency of measurements.
- Privacy-driven: This strategy targets to preserve end-user privacy by altering the accuracy of the individually retrieved results, without significantly altering their statistics. In particular, this strategy will (a) try to minimize the amount of times a specific sensor is requested for data (using also approaches developed for the data-driven and time-driven strategies), and (b) add "noise" to the retrieved results using differential privacy techniques.

The above strategies are implemented solely at the receiver-side, i.e. at the data collector (see Figure 1). This is the only option possible if the sensors cannot implement specific collection strategies, as is the case with the FIESTA-IoT platform which we consider in our experiments. Another advantage of the receiver-driven approach is that because sensor nodes typically have small processing and storage capabilities, the range of strategies they can implement can be limited. Unlike sensor nodes, the data collector typically has significantly more processing capabilities.

The architecture of our BeSmart framework is shown in Figure 1. The strategy layer, which implements the aforementioned four strategies, determines the time period at which measurements are requested from the FIESTA-IoT platform. The FIESTA-IoT platform manages IoT data from heterogeneous systems and environments and their entity resources (such as smart devices, sensors, and actuators), and was developed by the EU-funded Federated Interoperable Semantic IoT/cloud Testbeds and Applications project. FIESTA-IoT enables experimenters to use a single Application Program Interface (API) for executing experiments over multiple IoT testbeds that are federated in a testbed agnostic way.

Next, we discuss in more detail the four strategies implemented in our framework: data accuracy-driven, time-driven, energy-driven, and privacy-driven data collection. These strategies will be evaluated with measurements obtained from the FIESTA-IoT platform in Section III.

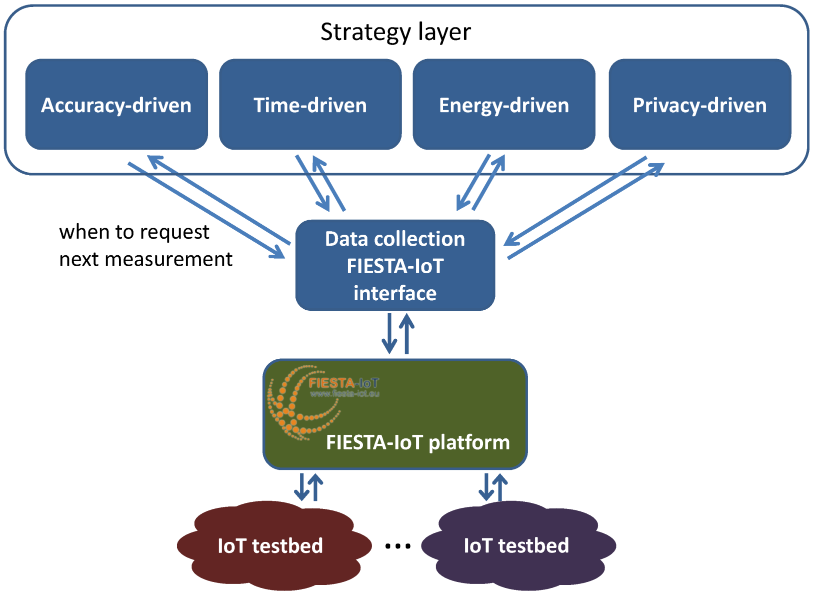


Fig. 1. BeSmart data collection architecture. The strategy layer determines the frequency of queries to the FIESTA-IoT platform.

A. Accuracy-driven data collection

The motivation for the accuracy-driven data collection strategy is that many applications require a particular data accuracy, and providing a higher accuracy offers no benefits. Hence, the goal of this strategy is not to select the period between measurement requests such that the time series of data measurements have the smallest deviation from the actual sensor values, which indeed can be the period at which the IoT sensor obtains measurements for a particular phenomena. Rather, the strategy seeks to reduce the frequency of the measurements, hence reduce the amount of resources (processing, communication, and storage) necessary for data collection, while achieving a target data accuracy.

The data accuracy-driven strategy, more specifically, seeks to achieve an accuracy such that the *last measurement* obtained differs from the *current sensor value* by at most a target accuracy; this target accuracy can be expressed as a simple percentage, e.g. the last measurement differs from the current sensor value by at most 10%. Of course, only an Oracle with knowledge of all the future sensor values can achieve the above goal 100% of the time. Figure 2 illustrates when measurements are requested by the Oracle, which has full knowledge of all future measurement values. A target data accuracy defines an accuracy interval corresponding to the last measurement that was obtained. A new measurement is requested whenever the current sensor value is outside the accuracy interval of the last measurement. Observe from Figure 2 that measurements are requested more frequently, i.e. the period between consecutive measurement requests is smaller, when the measured values change at a higher rate. We will use the Oracle as a benchmark to compare the performance of the proposed adaptation procedure that we describe next.

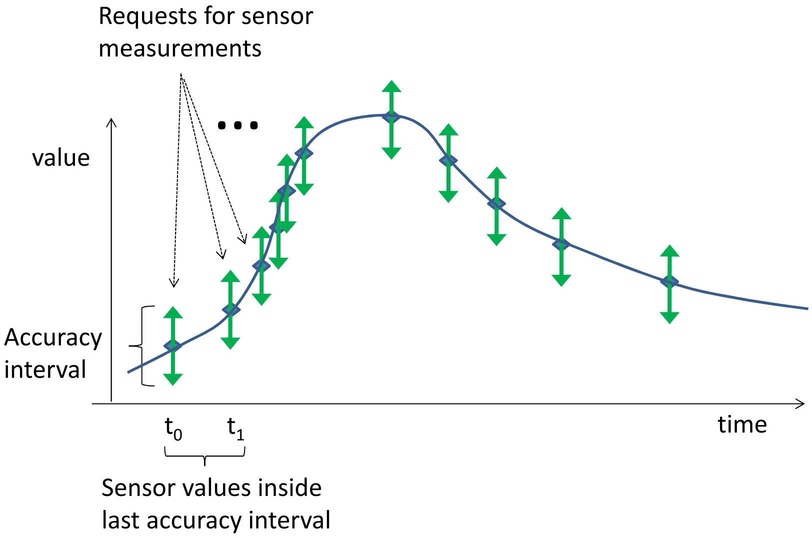

Requests for sensor measurements
value
Accuracy interval
t_0
t_1
time
Sensor values inside last accuracy interval

Fig. 2. Given a data accuracy target, which defines an accuracy interval, the goal is to request measurements such that at any time the current sensor value (continuous line in figure) falls within the accuracy interval of the last measurement. Only an Oracle can achieve this 100% of the time. The proposed procedure in Algorithm 1 tries to approximate this behavior.

Since the value of future measurements is not known, the Oracle cannot be applied in practice. Instead, we propose a procedure to adapt the time period between consecutive measurement requests based on the last measurement and the measurement when the period was last adapted, whose pseudocode is shown in Algorithm 1. Initially, the time period between measurement requests is equal to some small value, which can be the interval in which the sensor produces (or obtains) its measurements, that we will refer to as unit interval (line 9 in Algorithm 1). The time period T between consecutive measurements is increased by one unit interval (line 15), i.e. it increases linearly, if the current measurement is inside the accuracy interval based on the measured value when the interval T *was last adapted* (checked in line 15). If the current measurement is outside this accuracy interval, then T is reduced to half its value (line 17), i.e. it decreases multiplicatively.

The above additive increase and multiplicative decrease (AIMD) procedure resembles the AIMD adaptation of TCP's congestion window. However, the AIMD procedure is used to adapt a very different quantity: in TCP the AIMD procedure adapts the number of segments that can be sent without having received an acknowledgement. In our proposal the procedure adapts the time period between consecutive measurement requests. Moreover, the motivation for the additive

Algorithm 1 Procedure for adapting the time period between measurement requests

```
Variables:
$T$ : period between consecutive measurement requests
$i$: measurement request number
$k$: measurement request number when period was last modified
$M_i$: measured value returned for request $i$
$A \leftarrow q$: target accuracy, e.g. $q = 0.1$ for 10% accuracy
Algorithm:
$i \leftarrow 1$ and request first measurement $M_1$
$T \leftarrow 1$
$k \leftarrow 1$
loop
    Wait until time $T$ has elapsed since last measurement request
    $i \leftarrow i + 1$ and request new measurement $M_i$
    if $M_i \in [(1 - A) \cdot M_k, (1 + A) \cdot M_k]$ then
        $T \leftarrow T + 1$
    else
        $T \leftarrow T/2$
        $k \leftarrow i$
    end if
end loop
```

increase is that the procedure slowly tests larger periods between consecutive measurements, while the reduction to half is conservative to avoid requesting measurements too far apart, hence having measurements that deviate from the actual sensor values. A key feature of the proposed AIMD adaptation of the measurement period is that it is not based on some data model and involves no tuning parameters.

B. Time-driven data collection

The time-driven data collection strategy tries to ensure that the time a new measurement is received minus the timestamp of the last measurement is below some maximum delay. This elapsed time corresponds to the timeliness (or freshness) of data measurements. If there was no response delay from the time a measurement request is sent to the IoT measurement platform and the time that the response to the request containing the measurement is received, then it would be sufficient to send measurement requests with a period equal to the target maximum delay. In reality, if the time-driven strategy seeks to achieve a maximum delay T from the timestamp of the last measurement, then the next measurement request should be sent earlier than T by an amount that depends on the average and the distribution of the IoT measurement platform response times. This will be experimentally investigated in Section III.

C. Energy-driven data collection

The energy-driven strategy seeks to maximize the net bene?t, which is taken to be the difference between the utility that is achieved for some data accuracy minus the power consumption for the corresponding measurements that are necessary to achieve that data accuracy. The utility for a speci?c accuracy depends on the application requirements. Typically, applications can require a specific target data accuracy, and providing a higher accuracy offers no benefits. This results in the utility curve for accuracy higher than the specific target data accuracy being flat, as we discuss in Section III-C.

D. Privacy-driven data collection

This strategy targets to preserve end-user privacy by altering the accuracy of individual measurements but at the same time achieving "adequate accuracy" in the accumulated results. This is achieved by adding "noise" to the retrieved results using *differential privacy* [1].

The goal of differential privacy is to allow the extraction of statistics for a population of users without revealing any information about particular individuals. This is achieved with the addition of some "noise" to the statistics extraction process. The type and amount of noise added is such that the contribution of a single individual to the calculated statistics is not "significant." Hence, differential privacy guarantees that the output of the statistics calculation process is only slightly impacted by the contributions of a single individual. However, extracting information about a group of users in a privacy preserving way provides only partial security: a provider should still be trusted to collect and maintain user-provided sensitive information. This is not necessary if noise is added at the source during the data collection phase and not during the data processing phase. A trivial approach for implementing this functionality is the so-called "survey based on random responses" [2]. In a nutshell, with this approach, if a user is asked a question that can be answered with a "Yes" or "No," (for example, "Is your access network congested?"), then she flips a fair coin, in secret, and answers the truth if it comes up tails. Otherwise she flips another coin in secret and answers “Yes” if it comes up tails or “No” otherwise. This approach allows users to retain very strong deniability, while at the same time the real ratio of “Yes” answers can be accurately estimated using $2 \cdot (Y - 0.25)$, where Y is the portion of “Yes” responses.

For our privacy-driven strategy we use the RAPPOR protocol. The RAPPOR protocol [3] is an extension of the random responses approach, developed by Google and used in Google Chrome for collecting user statistics. This protocol extends the random response approach as follows: (i) it enables questions that can be asnwered with multiple responses, and (ii) it protects user privacy even if the same question is asked many times. The RAPPOR data collection process is extremely lightweight, hence it can be implemented even in constrained devices. The basic idea behind RAPPOR is that whenever a user is asked a question, she is provided with a set of possible responses. For each possible response the user plays a variation of the “random response” game. Eventually the user responds with a bit vector of size equal to the responses’ set: a bit in the vector set to 1 means that the corresponding response is selected by the user as one of the answers to the question. Due to the randomness of the response process, a user may select multiple responses or none of the responses. Furthermore, a selected response may or may not correspond to the user’s real answer.

The RAPPOR protocol requires an adequate number of responses in order to provide accurate results. In the context of our work, RAPPOR would require multiple sensor measurements to achieve average value calculation close to the “real” one. For this reason, RAPPOR is not ideal when high accuracy is required and there is a small number of sensor measurements (e.g., measuring the temperature of a single room using two temperature sensors). On the other hand, RAPPOR is tuneable and considers an accuracy/privacy trade-off. In particular, whereas in the simple random response game, a user flips a fair coin, in secret, and answers the truth if it comes up tails, using RAPPOR, bias can be added to this coin, hence the probability that a user answers the truth can be increased (better accuracy, less privacy) or decreased (better privacy, less accuracy).

Moreover, and as we discuss in the following section, the effectiveness of RAPPOR is affected by the number of possible answers a user can give to a question. So, in that case, it is up to the solution designer to decide the desired level of accuracy. For example, in this paper we want to measure the average temperature of an area. In order to minimize possible responses, we rounded sensor measurements to the closest integer, since each decimal digit increases 10 times the number of possible answers.

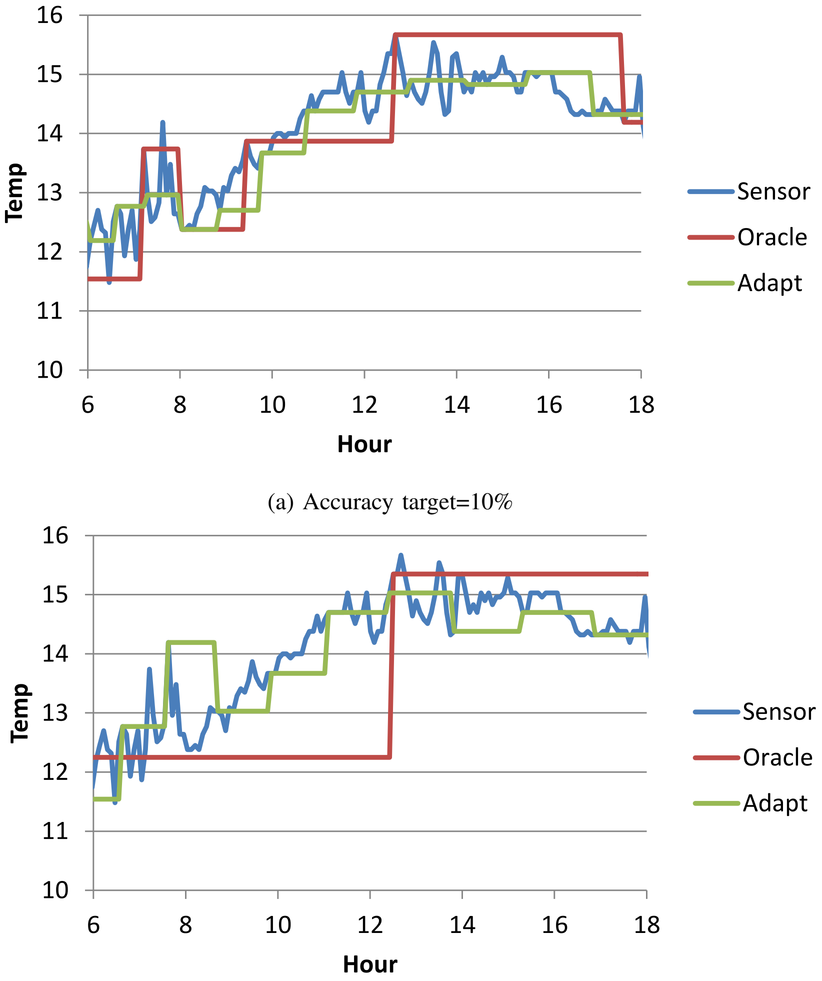
III. Experiments

In this section we present experiments for the data accuracy-driven and the privacy-driven data collection strategies. Our experiments consider measurements of different data types (phenomena), namely temperature, humidity, and ozone (O3) concentration, and three testbeds: SmartSantander, FINE, and Tera4Agri testbeds, whose measurements are obtained through the FIESTA-IoT platform. The SmartSantander testbed is a large-scale smart city deployment located in Santander, northern Spain. The FINE testbed is located in the city of Heraklion on the island of Crete, Greece. The Tera4Agri testbed is located on a farm in the Apulia region, Italy.

A. Accuracy-driven data collection

Figure 3 shows the temperature measurements for the Oracle, which has knowledge of the future temperature values, and the AIMD adaptation procedure, for two data accuracy targets: 10% and 20%. The Oracle requests measurements based on the knowledge of future temperature values as discussed in Section II-A and shown in Figure 2. At time periods where the graph for the Oracle and the AIMD adaptation procedure is horizontal, measurements are requested only at the beginning of the corresponding period. Hence, measurements are requested when the graph increases or decreases. Observe in Figure 3 that the measurements for the Oracle, compared to those for the AIMD adaptive scheme, are farther from the actual sensor values but still within the target accuracy interval; this is expected since the Oracle tries to reduce the number of measurements while allowing the maximum deviation of the measurements from the actual sensor values to be equal to the target data accuracy. Comparing Figure 3(a), which was obtained for target data accuracy 10%, with Figure 3(b), which was obtained for target data accuracy 20%, we observe that a larger data accuracy results in fewer measurements for both the Oracle and the AIMD adaptation procedure.

Figure 4(a) shows that the AIMD adaptation procedure can follow the varying trend of the temperature quite well, even though the values of the temperature exhibit periodicity and non-stationarity. Figure 4(b) shows the adaptation of the measurement period of the AIMD procedure, illustrating its additive increase and multiplicative decrease behavior. Finally, Figure 4(c) shows that the magnitude of the deviations for the AIMD procedure, i.e. the relative difference of the measured values compared to the sensor value when the measured values are outside the target accuracy, is typically less than 10%.



(a) Accuracy target=10%

(b) Accuracy target=20%

Fig. 3. Measurements with the Oracle and the AIMD adaptation procedure for two target accuracies: 10% and 20%. Comparison of the two graphs illustrates the tradeoff between the target data accuracy and the number of measurements.

Table I quantifies the gains in terms of the reduced number of measurements of the Oracle and the AIMD adaptation procedure, for a 96 hour time window that includes a total of 1147 temperature values with an interval of approximately 5 minutes between consecutive values. Both the Oracle and the AIMD adaptation procedure reduce the number of measurements by over 90%, i.e. more than 90% of the measurements produced by the sensor did not need to be collected, while still maintaining the target data accuracy. In particular, for 10% target data accuracy, with the AIMD measurement period adaptation procedure only 86 out of the total 1147 temperature values are requested (7.5%), which is very close to the number of measurements requested with the Oracle, which is 68 (5.9%). Moreover, the gains increase when the target

TABLE I

TEMPERATURE, SMARTSANTANDER, 96 HOUR TIME WINDOW, TOTAL # OF SENSOR VALUES: 1147

Target accuracy	Oracle	AIMD adaptation
10%	measurements: 68 (5.9%)	measurements: 86 (7.5%) deviations: 56 (4.9%)
20%	measurements: 9 (0.8%)	measurements: 48 (4.2%) deviations: 2 (0.05%)

data accuracy increases from 10% to 20%, i.e. the data accuracy becomes less stringent. The percentage of measurements which are outside the target data accuracy interval for the AIMD adaptation procedure is 4.9% and 0.05% of the total number of sensor values (1147), for a target data accuracy 10% and 20%, respectively; this means that, for a 10% target data accuracy, only 56 out of the 1147 temperature values (4.9%) were outside the 10% accuracy interval of the last measured value. For a 20% target data accuracy, only 2 out of the 1147 (0.05%) temperature values were outside the 20% accuracy interval. On the other hand, since it has knowledge of future temperature values, which of course in practice is not possible, the Oracle can request measurements such that they all differ from the actual sensor values by at most the target data accuracy.

Next we consider measurements for the same metric, temperature, obtained from the FINE and Tera4Agri testbeds. Table II quantifies the gains in terms of the reduced number of measurement requests for temperature measurements obtained from the second testbed, FINE, whose temperature sensors provided measurements every 10 minutes. Compared to the results in Table I, the number of measurements for both the Oracle and AIMD adaptation procedures is higher, but the gains are still significant: the AIMD procedure achieves over 80% reduction of the total number of measurements for a 10% target data accuracy, with only 5% deviations. The gains in terms of the reduced number of measurements are in terms of the highest rate at which the sensor produces measurements, which is different for different testbeds. Table III quantifies the gains for temperature measurements obtained from the third testbed, Tera4Agri, whose temperature sensors provided measurements every 17 minutes. Table III shows that the number of measurements for both the Oracle and AIMD adaptation procedures is higher than both the SmartSantander and FINE testbeds. This occurs because in the Tera4Agri testbed sensors provide temperature

TABLE II

TEMPERATURE, FINE, 96 HOUR TIME WINDOW, TOTAL # OF SENSOR VALUES: 628

Target accuracy	Oracle	AIMD adaptation
10%	measurements: 48 (7.6%)	measurements: 109 (17.4%) deviations: 32 (5.1%)
20%	measurements: 22 (3.5%)	measurements: 66 (10.5%) deviations: 29 (4.6%)

TABLE III

TEMPERATURE, TERA4AGRI, 144 HOUR TIME WINDOW, TOTAL # OF SENSOR VALUES: 500

Target accuracy	Oracle	AIMD adaptation
10%	measurements: 110 (22%)	measurements: 178 (35.6%) deviations: 74 (14.8%)
20%	measurements: 75 (15%)	measurements: 137 (27.4%) deviations: 61 (12.2%)

measurements every 17 minutes, compared to approximately 10 minutes for the FINE testbed and 5 minutes for the SmartSantander testbed. A larger interval between consecutive sensor measurements results in the temperature values of consecutive measurements differing more. The larger interval between sensor measurements for the Tera4Agri testbed (17 minutes) was also the reason we considered a larger 144 hour time window in Table III.

Table IV quantifies the gains in terms of the reduced number of humidity measurements. The AIMD procedure again achieves significant gains by reducing the number of measurements requested by more than 85%, compared to the case where all measurements produced by the sensor are requested. The results in Table IV also illustrate the tradeoff between data accuracy and number of measurements, for both the Oracle and the AIMD adaptive procedure. Observe that the difference in performance of the Oracle and the AIMD procedure is larger for the humidity measurements than for the temperature measurements, as can be seen in Table I; this is because the humidity changes more abruptly compared to the temperature for the time window in which the experiments were performed. A similar dependence, but due to the interval where sensors in different testbeds provided measurements, was observed above for the temperature.

TABLE IV

HUMIDITY, SMARTSANTANDER, 96 HOUR TIME WINDOW, TOTAL # OF SENSOR VALUES: 977

Target accuracy	Oracle	AIMD adaptation
10%	measurements: 34 (3.5%)	measurements: 134 (13.7%) deviations: 51 (5.2%)
20%	measurements: 11 (1.1%)	measurements: 56 (5.7%) deviations: 52 (5.3%)

TABLE V

O3, SMARTSANTANDER, 96 HOUR TIME WINDOW, TOTAL # OF SENSOR VALUES: 1063

Target accuracy	Oracle	AIMD adaptation
10%	measurements: 30 (2.8%)	measurements: 82 (7.7%) deviations: 53 (5%)
20%	measurements: 8 (0.8%)	measurements: 49 (4.6%) deviations: 12 (1.1%)

Table V quantifies the gains for ozone (O3) concentration measurements. This table shows that the reduction of the number of measurements achieved with the proposed AIMD procedure is significant: over 92% reduction of the measurements requested by the sensor for 10% accuracy and over 95% for 20% accuracy.

B. Time-driven data collection

The time-driven strategy tries to ensure that the elapsed time since the timestamp of the last measurement is below some maximum delay, which corresponds to the timeliness of data measurements. To achieve this maximum delay target, the time a new measurement is requested depends on the delay from the time a measurement request is sent to the FIESTA-IoT platform and the time that the response to the request is received. A sample of this response time is shown in Figure 5(a), while the corresponding cumulative distribution function (CDF) is shown in Figure 5(b). Although the average FIESTA-IoT response time is small, approximately equal to 0.47 seconds, the response time can result in values which are above 5 seconds; such values may impact time-critical applications, which motivate the importance of a time-driven data collection

TABLE VI

NUMBER OF MEASUREMENTS EXCEEDING TARGET DELAY ($T = 8$ S) AND AVERAGE INTERVAL OF MEASUREMENT REQUESTS FOR DELAY-DRIVEN STRATEGY

D	Delay violations	Average measurement interval (s)
Avg $+ 0.25 \cdot$ StdDev	107 (12.7%)	7.43
Avg $+ 0.5 \cdot$ StdDev	57 (6.8%)	7.33
Avg $+ 1.0 \cdot$ StdDev	43 (5.1%)	7.12
Avg $+ 1.5 \cdot$ StdDev	29 (3.5%)	6.91
Avg $+ 2.0 \cdot$ StdDev	23 (2.7%)	6.70

strategy. The delay values did not follow a known distribution, and for this reason below we perform a measurement-based investigation for the tradeoff between the number of measurements obtained with a specific measurement interval and the percentage of measurements that exceed the maximum target delay.

To achieve a target delay T from the timestamp of the last measurement until the time the next measurement is received, the request for the next measurement should be sent after time T minus some amount D that depends on the average and distribution of the FIESTA-IoT platform response times. However, due to the variability of the response times, we cannot ensure that all measurements are received with the target delay. Table VI shows the number of measurements that are received with a delay above the maximum target delay for different functions of D, namely $D =$ Avg $+ a \cdot$ StdDev for $a = 0.25, 0.5, 1,$ 1.5, and 2; where Avg and StdDev is the average and standard deviation of the delay measurements, respectively. The results were obtained with the delay measurements for the 3.5 hour window shown in Figure 5, which included a total of 840 delay values. The delay D is estimated using all delay values prior to a particular measurement. As expected, the second column in Table VI shows that if measurements are requested earlier, i.e. the value of a is higher, then the number of measurements that are received in violation of the target delay are fewer.

According to the above time-driven data strategy, the time interval of measurement requests can be adapted, based on the estimate of D. Figure 6 shows the interval of measurement requests, for a target delay of 8 seconds and $D =$ Avg $+ 1.5 \cdot$ StdDev. The figure shows that the time interval of measurement requests converges after a few hundred measurement requests.

The third column in Table VI shows the average interval of measurement requests for different estimations of D, when the target delay is 8 seconds. Comparison of the second and third columns in Table VI quantifies the tradeoff between the number of measurements that exceed the target delay and the interval of consecutive measurement requests, which determines the total number of measurements inside a time window and the corresponding energy consumption: requesting measurements at a smaller interval, 6.7 rather than 7.3 seconds (8% smaller), reduces the number of measurements exceeding the maximum target delay from 6.8% to 2.7%. Moreover, the relative gains from the reductions of delay violations versus the increase of the number of measurements is higher for smaller values of a: an increase of a from 0.25 to 0.50 results in a 1.3% increase of the number of measurements while the target delay violations are reduced by 47%. However, an increase of a from 1 to 2 results in the same reduction of the percentage of delay violations, but with a 6% increase of the number of measurements.

C. Energy-driven data collection

The energy-driven strategy seeks to maximize the net benefit, which is the difference between the utility for a specific data accuracy minus the power consumption for the measurements that are necessary to achieve this data accuracy. We consider the case of applications that require a target data accuracy and providing a higher data accuracy does not offer any benefits. The utility for such applications corresponds to the curve labelled "U(accuracy)" in Figures 7(a) and 7(b). Specifically, the utility in Figure 7(a) corresponds to an application which is indifferent to data accuracy values below 0.1 (i.e. 10% data accuracy). On the other hand, the utility for data accuracy above 0.1 decreases; hence, there is a "knee" at data accuracy equal to 0.1. Figure 7(b) corresponds to an application which is indifferent to data accuracy values below 0.2 (i.e. 20% data accuracy).

The energy cost in Figures 7(a) and 7(b) is determined by the number of measurements required for achieving the specific data accuracy. The energy cost curves were produced by running multiple data accuracy-driven experiments for accuracy values ranging from 0.05 to 0.3, with the data accuracy increasing in steps of 0.025. The only assumption made for the energy results is that the energy consumption is a linear function of the number of measurements.

The benefit in the figure is determined by the difference of the utility for different data accuracies minus the corresponding energy consumption for the number of measurements requested from the FIESTA-IoT platform, while taking into account the percentage of measurements that were inside the target data accuracy. Subtracting the cost from the utility to obtain the benefit would typically require converting cost units to utility units through some relative unit cost factor. Interestingly, the results presented in this section are independent of this relative cost factor. This is due to the fact that the “knee” of the utility curve is for data accuracy values where the energy cost increases slowly as the data accuracy becomes smaller. In general, this result occurs as long as the slope of the cost curve is smaller than the slope of the utility curve for data accuracy values close to the “knee” of the utility curve. Based on our measurements, this is always the case for the shape of the utility curve considered in this section and shown in Figures 7(a) and 7(b), and for data accuracy values above 5%, which are considered realistic.

The results in the figure show that, for the specific utility shape considered where the utility is flat up to some data accuracy value and then starts to decrease (i.e. there is a “knee” in the utility curve), the optimum target data accuracy is equal to the accuracy where the utility starts to decrease. Interestingly, Figures 7(a) and 7(b) show that this conclusion for the optimum target data accuracy is independent of the data accuracy value at which the utility starts to decrease, for the reason identified above. Moreover, this result is independent of the absolute amount of energy for each measurement, as long as the energy for each measurement is independent of the rate of measurements, which is equivalent to the assumption mentioned above that the energy consumption is a linear function of the number of measurements.

The tradeoff between data-accuracy and energy consumption is shown in Figure 8, for both the Oracle and the AIMD adaptive procedures. This figure shows that the energy gains increase quickly with increasing target accuracy (i.e. the target data accuracy becomes less tight) when the accuracy has small values. However, for data accuracy values higher than 0.15, the gains stabilize at a high value: more than 95% for the adaptive scheme, which is very close to the gains achieved with the Oracle (98%). It is important to note that, as above, the results in Figure 8 related to energy gains are independent of the absolute value of the energy consumption, and rely solely on the assumption that the consumption is a linear function of the number of measurements. Finally, we highlight the fundamental design difference of the data accuracy-driven strategies proposed in this paper, compared to other proposals targeting the highest achievable data accuracy that result in higher energy costs, due to the larger number of required measurements to achieve the higher accuracy, without a corresponding benefit for the application.

D. Privacy-driven data collection

The goal of this strategy is to maximize measurements accuracy while maintaining end-user privacy. In particular, we want to extract averages of measurements without learning the measurements of each individual sensor. Here we assume that these can be used for extracting sensitive information about a user. Although the types of measurements provided by the Fiesta-IoT testbeds do not fit in this category (and probably on purpose), there are use cases that consider similar types of measurements and at the same time impact end-user privacy. For example, a system that measures the energy consumption of the buildings of a smart city can be used for deducing when somebody is in his/her house.

For the evaluation of the privacy-driven strategy we implemented the *basic one-time RAPPOR* algorithm, described in [3], and applied it for calculating averages of temperature measurements performed by sensors in the Smart Santander testbed. We applied the algorithm over the measurements of 70 sensors in order to calculate hourly average temperatures. Each sensor was provided with a set of possible temperatures and for each element in the set the random responses game was executed. In particular, each sensor compared its measured temperature with each provided value and responded whether or not they match. We consider a *fair coin*, hence the sensor answers with probability 0.5. the truth, with probability 0.25 Yes, and with probability 0.25 No. Hence, each sensor could respond "Yes" to some/all/none temperatures and its responses may or may not include the real measurement. An aspect that affects the accuracy of the calculation is the size of the set of possible answers. In the following we compare two cases: (i) an oracle knows the lowest and highest measured temperature, and (ii) we rely on past measurements. In the first case, the set of possible answers is smaller, and it ranges from the smallest measurement to the greatest. In the second case, however, the set is grater in order to cover possible variations. In any case, we are not wishing to focus on how this set is constructed, but instead, we want to illustrate how the range of possible answers affects accuracy. Figure 9 illustrates[2] the distribution of measurements when the set of possible temperatures is confined to the actual minimum and maximum temperature, Figure 9(a), and when its boundaries are broader, Figure 9(b).

Since the sensor responses are randomized, every time an experiment is repeated the sensor responses (and hence the calculated average temperature) differ. Figure 10 illustrates the calcu-

[2] In this paper we include a subset of our results. Interested users are encouraged to use our live demo located at https://mm.aueb.gr/fiesta/privacy.php

lated average temperature for a specific hour[3] compared to the real average temperature, when the set of possible temperatures is confined to the actual minimum and maximum temperature and when its boundaries are broader. The graph shows that the calculated averages are very close to the real averages, despite the addition of noise.

IV. Related work

Prior work on data collection in wireless sensor networks has proposed schemes that trade the quality of the data collected for energy efficiency. This work includes schemes based on models at the data collector that capture the correlation of data to reduce the number of data queries to sensors [4], [5], [6]. Some proposals combine pull-based (located at the data collector) and push-based (located at the sensors) mechanisms for reducing the number of messages [7], [8], [6] or consider purely push-based update strategies [9]. The data collection strategies investigated in this paper differ from the above in that they do not rely on models, as in [5], [4], [6], but rather adapt the period between consecutive measurements based on whether the measurements lie inside a target data accuracy interval; this is a fundamental difference that makes the strategies proposed in this paper application-aware, since they take into account the requirements of applications in terms of their target data accuracy. Moreover, the proposed data collection procedures are implemented solely at the data collector side (pull-based), hence do not require mechanisms at the sensor side, which is not possible when the sensor testbeds are not directly accessible or are under different administrative control, as is the case of the testbeds federated under the FIESTA-IoT platform.

In addition to the temporal correlation of data measurements, the spatial correlation of measurements from sensors located in the same area can be exploited, as in [10], [11], [12], which investigate in-network mechanisms for quality-driven sensor cluster construction and estimation of probabilistic models for capturing the temporal and spatial correlation. As noted above, our approach differs in that we consider pure pull-based procedures. The work in [13] proposes a utility-based model for optimally assigning data queries to sensors in a participatory sensing system, while [14] investigates a quality-driven function to select a fixed number of sensors for accomplishing a sensing task.

The trade-off between privacy and accuracy is a well studied problem which still attracts researchers' attention; see for example [15], [16], [17]. These works try to add as much noise

[3] The graphs for all other hours follow a similar pattern.

as possible to the individual sensor measurements, while influencing as little as possible the aggregate statistics. The contribution of this paper is to evaluate the efficiency of these techniques using real IoT measurements, which is a different domain than the one differential privacy has been applied up to now.

V. Conclusions and future work

The proposed accuracy-driven data collection procedure employs an additive increase / multiplicative decrease (AIMD) adaptation of the time period between consecutive measurements. Experiments have shown that such an AIMD procedure is highly robust for measurements of different data types (phenomena), namely temperature, humidity, and ozone, and for measurements from three different testbeds. The AIMD adaptation procedure is able to track the trends of the values measured in the presence of periodicity and non-stationarity. Moreover, unlike other data collection schemes, it has no tuning parameters. The experiments for the privacy-driven strategy have shown that, despite the small number of sensors, the addition of noise to the sensor measurements using differential privacy has a negligible effect on the aggregate statistics. Ongoing work is investigating the tradeoff between data accuracy and power consumption. Additionally, we are combining the adaptation of the measurement period (temporal adaptation) with the dynamic selection of the subset of sensors to request measurements from (spatial adaptation); such an approach can yield higher energy efficiency, without sacrificing data accuracy.

Acknowledgements

This work was supported in part by the EU FIESTA-IoT project (Grant Agreement number: 643943 / H2020-ICT-2014-1) through the Open Call 3 Experiment “BeSmart: Smart IoT Data Collection” and by the Athens University of Economics and Business Research Center.

References

[1] C. Dwork, F. McSherry, and A. Nissim, Kobbiand Smith, “Calibrating noise to sensitivity in private data analysis,” in *Theory of Cryptography*, S. Halevi and T. Rabin, Eds. Springer Berlin Heidelberg, 2006, pp. 265–284.

[2] S. L. Warner, “Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias,” *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.

[3] U. Erlingsson, V. Pihur, and A. Korolova, “RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response,” in *Proc. of ACM SIGSAC Conference on Computer and Communications Security*, 2014.

[4] D. Chu, A. Deshpande, J. Hellerstein, and W. Hong, “Approximate data collection in sensor networks using probabilistic models,” in *Proc. of IEEE ICDE*, 2006.

[5] A. Deshpande, C. Guestrin, S. Madden, J. Hellerstein, and W. Hong, “Model-driven data acquisition in sensor networks,” in *Proc. of VLDB*, 2004.

[6] A. Jain, E. Chang, and Y.-F. Wang, “Adaptive stream resource management using Kalman Filters,” in *Proc. of ACM SIGMOD*, 2004.

[7] Q. Han, S. Mehrotra, and N. Venkatasubramanian, “Energy efficient data collection in distributed sensor environments,” in *Proc. of IEEE ICDCS*, 2004.

[8] Q. Han, D. Hakkarinen, P. Boonma, and J. Suzuki, “Quality-aware sensor data collection,” *International Journal of Sensor Networks*, vol. 7, no. 3, pp. 127–140, May 2010.

[9] X. Tang and J. Xu, “Adaptive Data Collection Strategies for Lifetime-Constrained Wireless Sensor Networks,” *IEEE Trans. on Parallel and Distributed Systems*, vol. 19, no. 6, pp. 721–734, June 2008.

[10] B. Gedik, L. Liu, and P. S. Yu, “ASAP: An Adaptive Sampling Approach to Data Collection in Sensor Networks,” *IEEE Trans. on Parallel and Distributed Systems*, vol. 18, no. 12, pp. 1766–1783, December 2007.

[11] C. Liu, K. Wu, and J. Pei, “An Energy-Efficient Data Collection Framework for Wireless Sensor Networks by Exploiting Spatiotemporal Correlation,” *IEEE Trans. on Parallel and Distributed Systems*, vol. 18, no. 7, pp. 1010–1023, 2007.

[12] C. Wang, H. Ma, Y. He, and S. Xiong, “Adaptive Approximate Data Collection for Wireless Sensor Networks,” *IEEE Trans. on Parallel and Distributed Systems*, vol. 23, no. 6, pp. 1004–1016, June 2012.

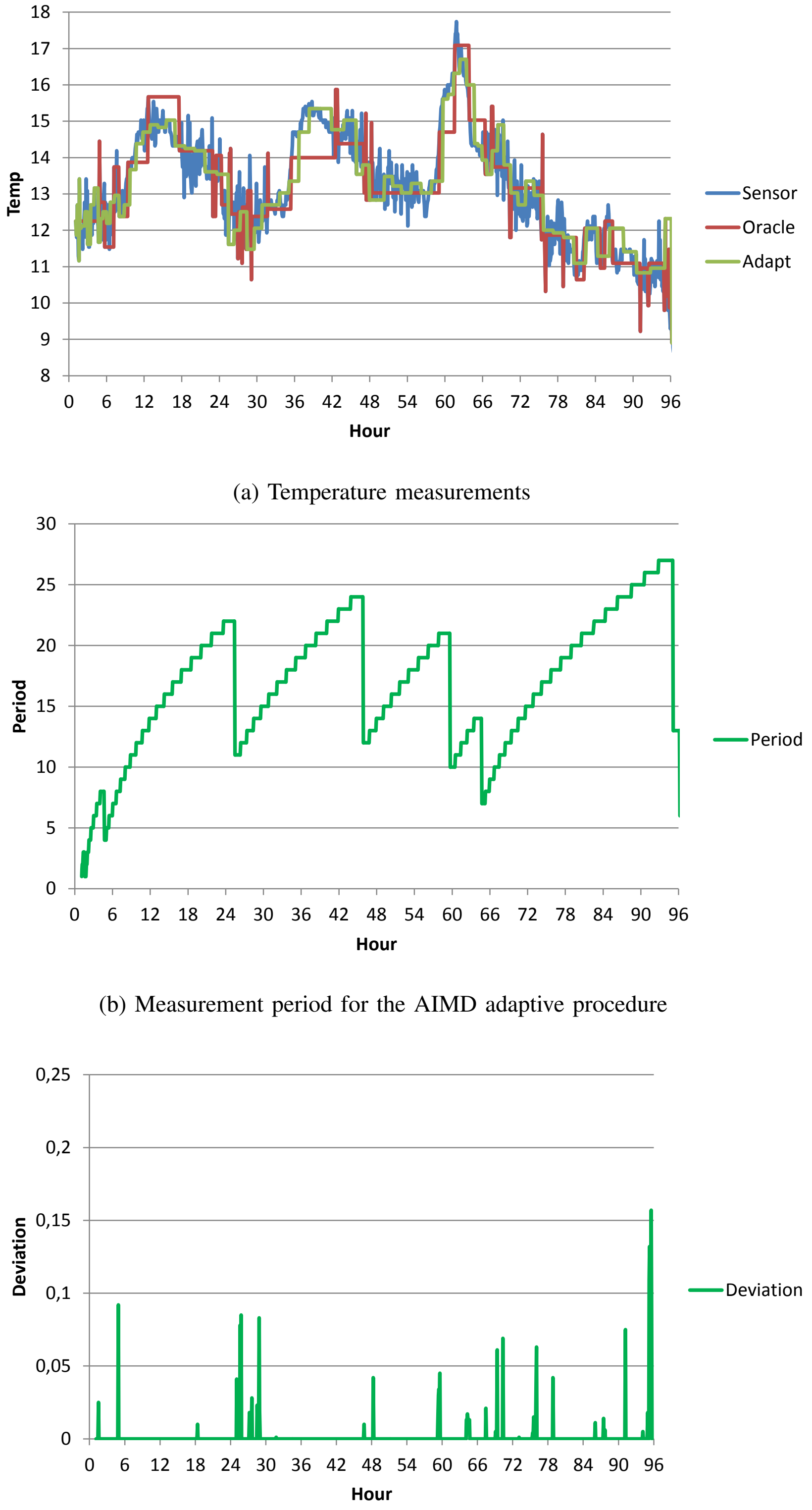
[13] M. Riahi, T. Papaioannou, I. Trummer, and K. Aberer, “Utility-driven data acquisition in participatory sensing,” in *Proc. of Int. Conf. on Extending Database Techn. (EDBT)*, 2013.

[14] M. Marjanovic, L. Skorin-Kapov, K. Pripuzic, A. Antonic, and I. Zarko, “Energy-aware and quality-driven sensor management for green mobile crowd sensing,” *Journal of Network and Computer Applications*, vol. 59, pp. 95–108, January 2016.

[15] M. Andres, N. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, “Geo-indistinguishability: Differential privacy for location-based systems,” in *Proc. of ACM SIGSAC Conference on Computer and Communications Security*, 2013.

[16] G. A. Fink, “Differentially private distributed sensing,” in *Proc. of 3rd IEEE World Forum on Internet of Things (WF-IoT)*, 2016.

[17] J. Chen, M. Huadong, and Z. Dong, “Private data aggregation with integrity assurance and fault tolerance for mobile crowd-sensing,” *Wireless Networks*, vol. 23, no. 1, pp. 131–144, 2017.

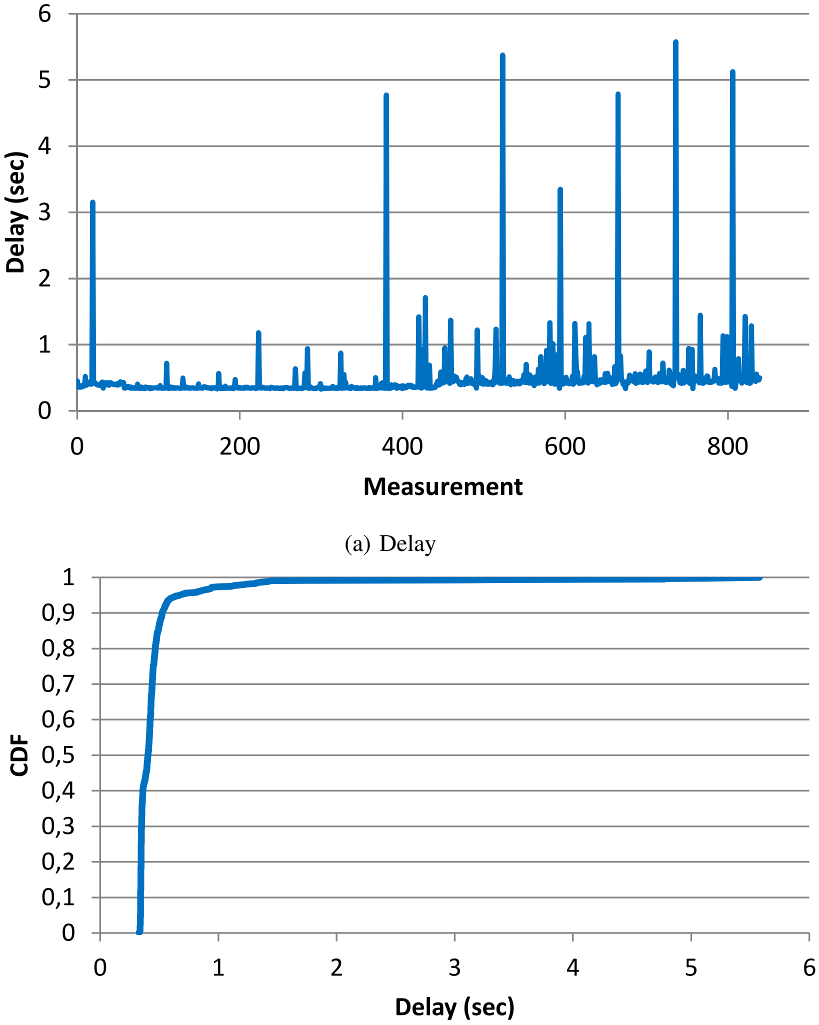


(a) Temperature measurements

(b) Measurement period for the AIMD adaptive procedure

(c) Deviations for adaptive scheme

Fig. 4. Performance of the AIMD adaptation procedure in a 96 hour time window of temperature measurements from SmartSantander, where sensors provide measurements every 5 minutes. Accuracy target=10%



(a) Delay

(b) Cumulative Distribution Function

Fig. 5. Delay to receive response from the FIESTA-IoT platform. Although the average FIESTA-IoT response time is small, approximately equal to 0.47 seconds, the response time can result in values which are above 5 seconds.

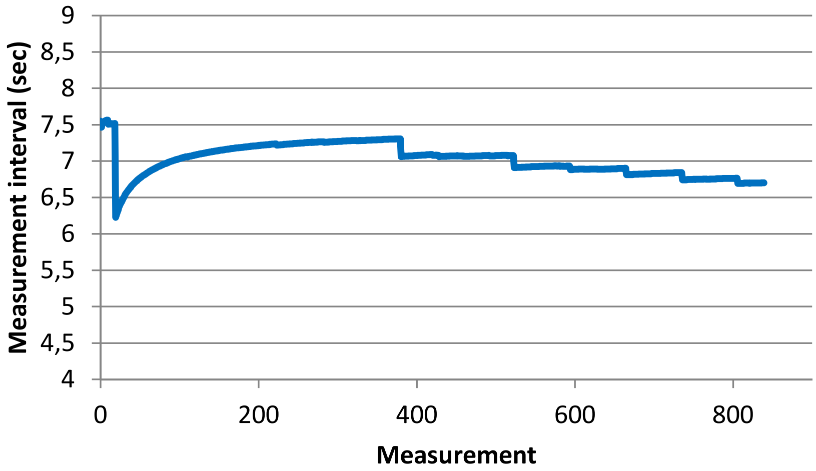
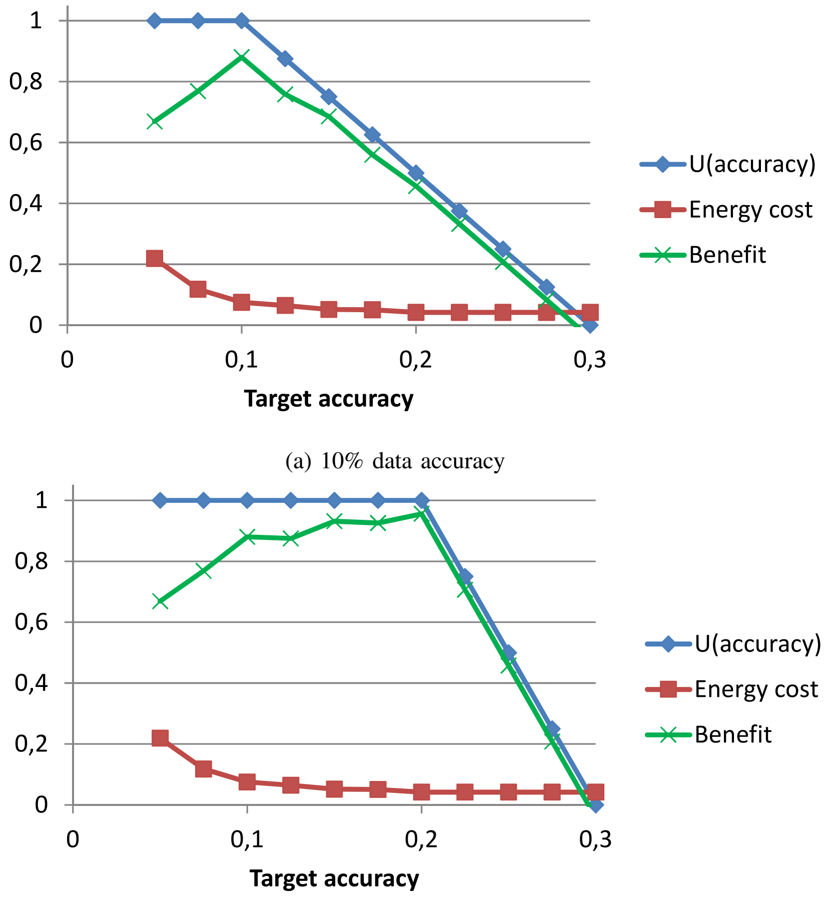


Fig. 6. Time interval of measurement requests for the time-driven strategy. The time interval at which measurements are requested converges after a few hundred measurements.



(a) 10% data accuracy

(b) 20% data accuracy

Fig. 7. Utility, energy, and net benefit as a function of target data accuracy.

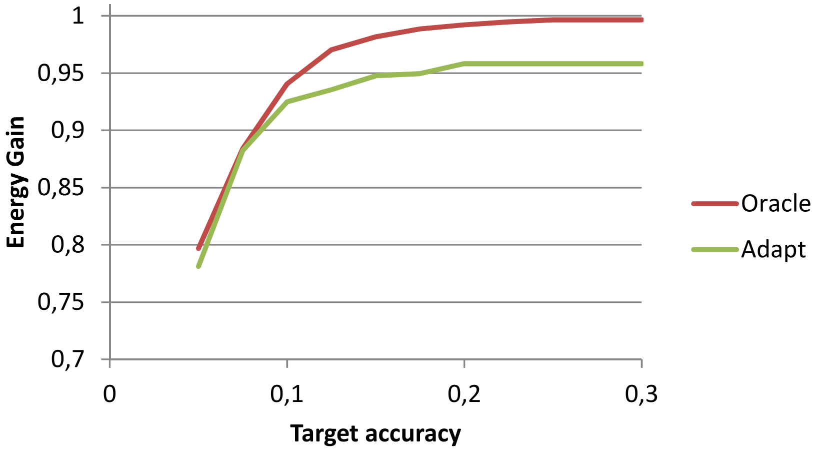
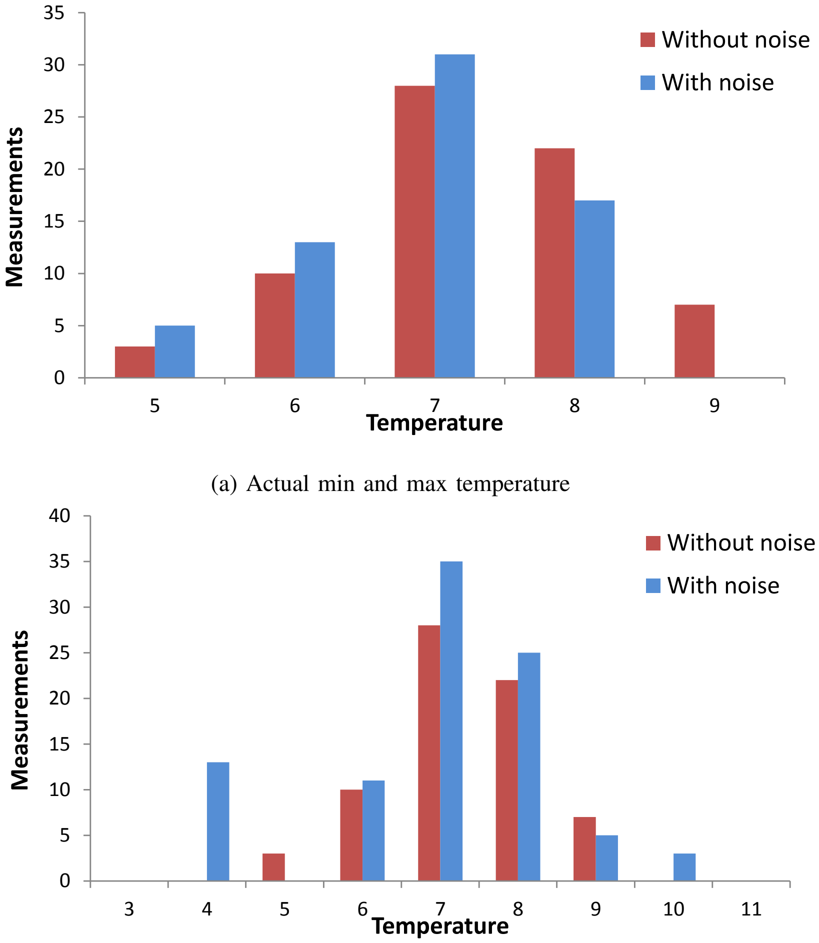


Fig. 8. Energy gains for different target data accuracy values.



(a) Actual min and max temperature

(b) Expanded min and max temperature

Fig. 9. Distribution of measurements with and without noise addition with the limits of the possible temperatures (a) equal to the actual min and max temperature, (b) expanded.

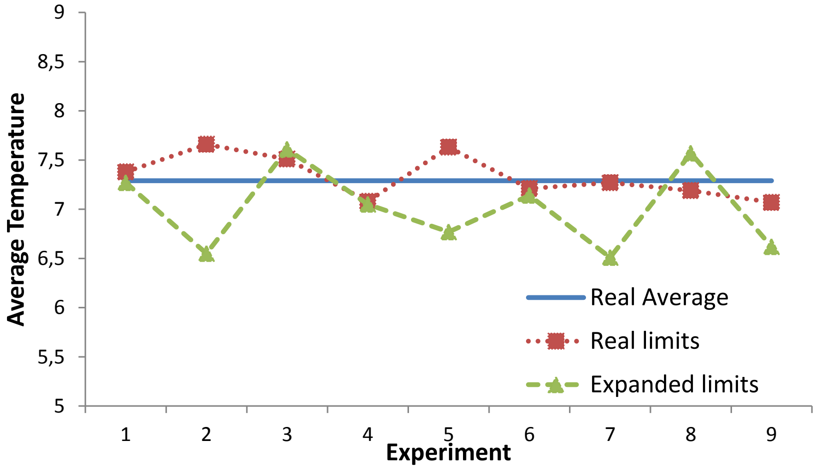


Fig. 10. Calculated average temperature per experiment.