Stream Correlation Monitoring for Uncertainty-Aware Data Processing Systems

Aleka Seliniotaki[1,2], George Tzagkarakis[1], Vassilis Christofides[1,2] and Panagiotis Tsakalides[1,2]

[1] Institute of Computer Science, Foundation for Research and Technology – Hellas
[2] Department of Computer Science, University of Crete

e-mail: {aseliniotaki, gtzag}@csd.uoc.gr, {christop, tsakalid}@ics.forth.gr

***Abstract*— In several industrial applications, monitoring large-scale infrastructures in order to provide notifications for abnormal behavior is of high significance. For this purpose, the deployment of large-scale sensor networks is the current trend. However, this results in handling vast amounts of low-level, and often unreliable, data, while an efficient and real-time data manipulation is a strong demand. In this paper, we propose an uncertainty-aware data management system capable of monitoring pairwise correlations of large sensor data streams in real-time. An efficient similarity function based on the truncated DFT is employed instead of the typical correlation coefficient to monitor dynamic phenomena for timely alerting notifications, and to guarantee the validity of detected extreme events. Experimental evaluation with a set of real data recorded by distinct sensors in an industrial water desalination plant reveals a high performance of our proposed approach in terms of achieving significantly reduced execution times, along with increased accuracy in detecting highly correlated pairs of sensor data streams.**

Keywords*— *Data management systems, online correlation monitoring, uncertainty quantification, wireless sensor networks

I. INTRODUCTION

During the last years, a significant technological progress is apparent in the design of large-scale self-organized wireless sensor networks (WSN) for carrying out various tasks, such as environmental monitoring and surveillance, in several sectors. The HYDROBIONETS project[1] is a characteristic example of such an infrastructure for water resource management. Specifically, it targets at developing a real-time micro-biological wireless networked control system for water desalination and treatment plants, providing the fundamental design principles of a wireless BioMEM network (WBN) with distributed multi-sensing and multi-actuation capabilities.

The HYDROBIONETS infrastructure project focuses on the monitoring of the complete water cycle in large-scale water treatment and desalination plants. The deployment of a WBN aims at monitoring critical microbiological and electrochemical parameters of water at different stages of the desalination process. This distributed, autonomous sensing is exploited by a series of functionalities, such as the detection of high fouling concentration in seawater, the control of biocide and chlorine dosage by measuring bacteria in seawater at different stages of water treatment (pre-filtered, pre-treatment and reverse osmosis phases) at periodic time intervals. These functionalities essentially provide the building blocks of the actuation process for water desalination at different locations in the plant.

The HYDROBIONETS data processing subsystem comprises of collaborating computational nodes, which observe and control distinct physical entities and dynamic phenomena. Rather than single stream statistics, such as average and standard deviation, data analysis is focusing on finding high *correlations* among pairs of data streams from distinct sensors. For instance, temperature and pressure sensors, which monitor an industrial plant could provide evidence of an increasing bacteria presence. Depending on their physical location in the plant we expect that corresponding data streams will be highly correlated.

More generally, a desalination plant operator may rely on such stream correlation engine to reveal interrelations between seemingly independent physical quantities monitored by distinct sensors, or to guarantee the validity of a detected extreme event (e.g. high chlorine concentration in the water) and provide the necessary notifications. . Moreover, in HYDROBIONETS, measurements from heterogeneous sensors, distributed over a geographic area, need to be processed efficiently in order to reconstruct the spatio-temporal behavior of desired physical variables or to detect, identify and localize sources and events of interest. Whereas traditional statistical machine learning provides well-established mathematical tools for data analysis [1][2][3][8], their performance is limited when processing high- dimensional data streams.

Furthermore, existing techniques for monitoring correlations exhibit several drawbacks. Mo re specifically, [5] studies the problem of maintaining data stream statistics over sliding windows, with the focus being only on single stream statistics. On the other hand, [6] introduced an extension for monitoring the statistics of multiple data streams, but the computation of correlated aggregates is limited to a small number of streams to be monitored. StatStream [4] is a data stream monitoring system, which enables the computation of single- and multiple-stream statistics. Its ability to achieve online monitoring of time synchronized streams is based on the combination of discrete Fourier transforms (DFT) with an appropriate hash technique. However, the main drawback of this technique is the difficulty to determine the hash function,

[1] http://www.hydrobionets.eu

This work is supported by HYDROBIONETS project (ICT-2011-7) funded by the European Commission in FP7 (GA-2011-287613).

which places the streams with similar behavior in neighboring cells of a grid structure. It is difficult to define an appropriate hash function for our data streams, since the streams we manage describe dynamic phenomena and their distribution is not known a priori.

In this paper, we overcome the limitations of previous approaches by introducing a computationally efficient similarity function, which enables the monitoring of pairwise correlations between high-dimensional sensor data streams on the fly. We note here that time synchronization is also performed between the acquired data streams, prior to the extraction of highly correlated pairs, based on their corresponding time stamps, which are available as a part of the transmitted packets. In particular, instead of computing all pairwise correlations between the original full-dimensional data streams, we exploit the compression property of the discrete Fourier transform (DFT) to concentrate the inherent energy content of a given signal in the first few high-amplitude coefficients, as in [4]. Then, a suitable peak similarity measure is applied on the associated pairs of truncated DFTs as a proxy for the corresponding correlation coefficients. Thus the problem of identifying highly correlated pairs of data streams is reduced to a problem of identifying pairs of truncated DFTs with high peak similarity values.

It is worth also to stress that usually WSN nodes do not handle any quality aspect of physical device data, but rather interface with a high-level representation and reconstruction of the sensed physical world. As a result, the HYDROBIONETS data processing subsystem has to additionally cope with data *uncertain* , where stream data may be incomplete, imprecise, and even misleading [7], thus impeding the task of an accurate and reliable decision making. *Uncertainty-aware data management* [9] presents numerous challenges in terms of collecting, modeling, representing, querying, indexing and mining the data. Since many of these issues are interrelated and cannot easily be addressed independently. Uncertainty has been recently recognized as an additional source of information that could be valuable during data analysis and thus should be preserved.

Another major functionality assigned to our data management system is to perform high-level operations, as the notification of *extreme events* from raw sensor data [2]. Since the detection of abnormal behavior is affected by the underlying uncertainty, incorporation of the estimated underlying uncertainty for the extraction of potential correlation between pairs of data streams is expected to yield more meaningful results. More specifically, a spreadsheet-based approach is employed in our uncertainty-aware data management system to identify, quantify, and combine the individual uncertainties corresponding to the most significant sources of uncertainty.

The performance of our proposed approach was evaluated on a set of real data provided by ACCIONA Agua, recorded by a set of distinct electromechanical sensors in La Tordera's desalination plant [2] . The results revealed a significant improvement of our method, in terms of highly reduced execution times and accurate estimation of the highly-correlated pairs of sensor streams, when compared with the typical correlation coefficient.

The remainder of the paper is organized as follows: Section II describes the application scenario, which motivates our proposed approach. Section III analyzes the algorithm for discovering data streams with a correlation above a specific threshold in a fast online fashion, while Section IV describes the method for identification and quantification of the underlying uncertainty. An experimental evaluation is carried out in Section V, while Section VI concludes and discusses possible future extensions.

II. PROBLEM DESCRIPTION

A. Monitoring Setting

The proposed data processing subsystem aims to support the HYDROBIONETS WSN infrastructure for multi-sensing and multi-actuation in water treatment and desalination plants. In our case, a desalination pilot plant is located in La Tordera, which is equipped with a number of various electrochemical sensors, scattered in distinct locations, for monitoring several physical and mechanical variables in the plant.

In order to perform timely actuation and provide guarantees for the validity of a detected extreme event, we need to monitor continuously and online the correlations between a number distinct data streams produced by sensors at different stages of water treatment (pre-filtered, pre-treatment and reverse osmosis phases), as well as their inherent uncertainty.

Depending on the type, the available electrochemical sensors may report a measurement within a predefined period of time, usually in the scale of a few seconds or minutes. Data processing of raw data streams is performed on the basis of *sliding windows*. In particular, a sliding window of recent measurement values is maintained, while the window moves with a predetermined step size when new measurements become available.

B. Monitoring Stream Correlations

Depending on the monitored phenomenon and the environmental conditions, the behavior of the data streams may significantly evolve over time. Changes in data characteristics (e.g. in data distribution) may indicate anomalies in the behavior of the monitored (aka an *extreme events*) or alterations in the data acquisition or transmission process. Detecting these behavior variations, is crucial for decision-making in water treatment or desalination since an accurate and timely detection of abnormal changes in sensor streams will enable a just-in-time actuation with the subsequent operational and maintenance cost savings. The most commonly used method to track the evolution of an observed variable is to compute the correlation coefficient among all pairs of sensor data streams.

III. DATA STREAM CORRELATION FRAMEWORK

Distinguishing efficiently between occasional and extreme events constitutes a major issue in monitoring and data

[2] http://aca-web.gencat.cat/aca/documents/ca/sensibilitzacio/desal_Tordera/dessalinitzacio_en.pdf

management systems. It is of great importance to ensure in real time, especially when we deal with massive data sets, that a true extreme event occurs and not some coincidence or system/network failure. The correlation between two or more sensor data streams characterizes their relationships and dependencies. Thus, the identification of highly correlated streams can be exploited as a further guarantee to verify the existence of a detected extreme event.

To clarify this point, consider the case of two data streams recorded by a pressure and a temperature sensor, respectively. When the two sensors are placed nearby, we expect that a high pressure is associated with an increased temperature, which means that the correlation of these two streams should be relatively high. Thus, we assume that a potential notification for an extreme temperature is indeed true, whether the measured pressure is also high. The only ambiguous point here is related to the determination of "high correlation". The degree of "high correlation" is related to the specific application and the end-user, who has the flexibility to define how much strictly or springily this degree will be.

Doing so, the set of available sensors is divided into subsets of highly correlated sensors. This clustering enables a more convenient and meaningful monitoring of the overall infrastructure by a system operator, who focuses only on a subset of sensors, where an abnormal behavior has been detected for at least one of its members.

In our proposed data management system, we implement a computationally efficient method for real-time extraction of highly correlated data streams by combining the DFT over sliding windows with a proper *peak similarity measure*. In order to account for the underlying uncertainty or other data ambiguities, as it will be introduced in the subsequent section, we restrict ourselves on the detection of pair of streams whose correlation is above a specific threshold.

A. *Fast online pairwise correlation estimation*

In our proposed approach, we search for sensor pairs whose correlation is above a predefined threshold $t_{threshold}$, in a fixed-sized sliding window. More specifically, let $\boldsymbol{s}$ be the reference stream, and ($\boldsymbol{y}_1$,$\boldsymbol{y}_2$,..,$\boldsymbol{y}_C$) be the set of streams with which we compute the pairwise correlations. For a predetermined correlation threshold $t_{threshold}$, the output of the process will be a subset of streams $\boldsymbol{y}_c$, for which the correlation with $\boldsymbol{s}$ is above $t_{threshold}$.

As a first step, each data stream values in the current window of length w, x_0,x_1,…x_{w-1}, are normalized to mean zero and variance one,

$$\hat{x}_i = \frac{x_i - \bar{x}}{\sigma_\chi} \quad (1) \quad \text{where} \quad \sigma_x = \sqrt{\sum_{i=1}^{w} (x_i - \bar{x})^2} \qquad (2)$$

and $\bar{x}$ denotes the mean value of $\boldsymbol{x}$. As a second step, the corresponding DFT of the normalized windowed data is computed. The DFT of a given data stream $\boldsymbol{x} = x_0, x_1, \ldots, x_{w-1}$ yields a set of complex numbers, the Fourier coefficients, $\boldsymbol{X} = X_0, X_1, .., X_{w-1}$, which are given by

$$X_f = \frac{1}{w} \sum_{k=0}^{w-1} x_k e^{-j2\pi fk/w} \qquad f = 0,1,..., w-1 \qquad (3)$$

A main property of DFT, which is exploited to reduce the computational cost, is the data compression capability. More specifically, the greatest portion of the signal's energy is concentrated in the first few high-amplitude Fourier coefficients, thus the original data can be approximated by a highly reduced set of coefficients.

The final step towards our real-time extraction of highly correlated sensor data streams is to identify those pairs $(\boldsymbol{s}, \boldsymbol{y}_c)$ with correlation above the given threshold $t_{threshold}$. In order to avoid computing the correlation between all pairs of streams $(\boldsymbol{s}, \boldsymbol{y}_c)$, we reduce the set of candidate streams only to those streams that will be highly correlated with $\boldsymbol{s}$ with high probability.

For this purpose, we introduce *peak similarity*, p_{sim}, as an appropriate similarity measure. More specifically, the similarity between two data streams $\boldsymbol{s}$, $\boldsymbol{y}$ is computed by employing a truncated set of the first n high-amplitude DFT coefficients, where $n << N$, and the peak similarity measure is defined as follows:

$$p_{sim}(s, y) = \frac{1}{n} \sum_{i=1}^{n} \left[1 - \frac{|S_i - Y_i|}{2 \cdot \max(|S_i|, |Y_i|)} \right] \qquad (4)$$

In order to account for the potential loss of information caused by the truncation of the set of DFT coefficients, the peak similarity measure does not employ the same threshold $t_{threshold}$ for finding the "similar" streams. Instead, we determine a new threshold t_{new}, with our proposed method reporting as "highly-correlated" pairs those streams $\boldsymbol{s}$ and $\boldsymbol{y}_c$ for which $p_{sim}(\boldsymbol{s},\boldsymbol{y}_c) > t_{new}$. However, special attention should be given on the selection of the threshold value t_{new}. From our experimental evaluation, employing data from a set of various distinct sensors, we observed that if we choose an "elastic" enough threshold t_{new}, then the subset of streams $\boldsymbol{y}_c$ with the highest peak similarity with $\boldsymbol{s}$ will also contain the highly correlated streams with $\boldsymbol{s}$ (that is, those with correlation coefficient above $t_{threshold}$). In our implementation we set $t_{new} = t_{threshold} + e$, where e is a small positive number (in our experimental evaluation described in the section V we set $e < 0.05$).

IV. MANAGING UNCERTAINTY IN DATA STREAMS

Having obtained the raw sensor data, our data management system estimates their underlying uncertainty. This process is performed in two consecutive phases, namely, identification of all the potential sources of uncertainty, followed by their quantification and propagation.

A. *Step 1: Identification of uncertainty sources*

Identification of uncertainty sources comprises the first step towards the design of an integrated uncertainty-aware data management system. In practice, the underlying uncertainty may arise due to several distinct sources, such as incomplete definition of the observed quantities, sampling effects and interferences, varying environmental conditions, and inherent uncertainties of the equipment.

A very convenient way of determining the uncertainty sources, together with their relations to each other, is to exploit the so-called *cause and effect (or Ishikawa) diagram*.

This diagram also ensures comprehensive coverage, while helping to avoid double counting of sources. Once the set of uncertainty sources is formed, their effects on the final result can be usually represented in terms of a measurement model.

As a typical example, Fig. 1 shows a cause and effect diagram for a temperature sensor. The first source of uncertainty is the sensor's functionality by itself. However, its performance is affected by several distinct factors, such as, its sensitivity and precision, its calibration, the operating temperature, and the water flow-rate and pressure. On the other hand, the accuracy of the recorded values depends also on the sensors deployment density and location, as well as on the sampling process we use. Possible misplacement or a very sparse time-sampling is expected to increase the uncertainty.

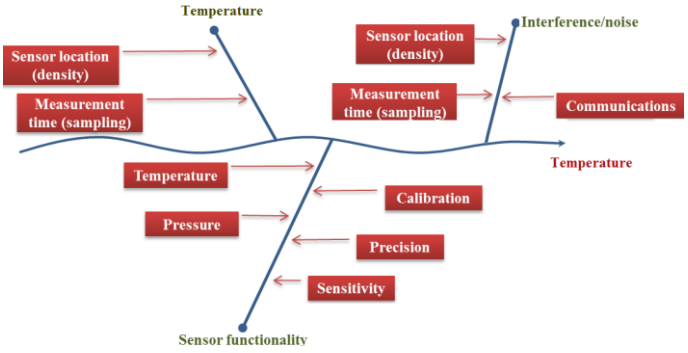

Figure 1: Cause and effect diagram for a temperature sensor.

B. Step 2: Quantification of uncertainty

The identification of uncertainty sources is followed by its subsequent quantification. This is done by estimating the uncertainty of each individual source and then combining them to obtain an overall uncertainty estimate. Towards asserting the uncertainty in raw data streams generated from our sensors, we recall here its distinction into two separate categories, namely, type A (or aleatory) and type B (or epistemic) uncertainty.

Uncertainties of type A are characterized by the estimated variances σ_i^2 (or the standard deviations σ_i), which are obtained by statistical analysis of the observations in the raw data streams. Following a sliding window approach, as mentioned in Section II.B, the variance σ_i^2 of the i-th sensor is estimated from its measurements in the current window. This is equivalent to obtaining a standard uncertainty from a probability density function (pdf) derived from an observed frequency (empirical) distribution. Let $\boldsymbol{y}$ be a data stream with N values $\{y_1,\ldots y_N\}$, which corresponds to a specific observed variable. Then, the standard uncertainty of $\boldsymbol{y}$, which is denoted by $u(y)$, is estimated by means of the corresponding standard deviation σ_y, estimated from the observations y_i, as follows,

$$u(y) = \frac{\sigma_y}{\sqrt{N}} \tag{5}$$

On the other hand, for uncertainties of type B, the estimated "variance" s_j^2 is obtained from an assumed probability density function based on our prior knowledge for the corresponding source of uncertainty, which may include a) previous measurement data, b) experience or knowledge of the properties of instrumentation and materials used, c) manufacturer's specifications, and d) calibration data. In general, concerning type B uncertainties, the quantification is performed either by means of an external information source, or from an assumed distribution. Typical assumptions for the prior distributions include the Gaussian (for instance, when an estimate is made from repeated observations of a randomly varying process, or when the uncertainty is given as a standard deviation or a confidence interval), the uniform (when a manufacturer's specification, or some other certificate, give limits without specifying a confidence level and without any further knowledge of the distribution's shape), and the triangular distribution (when the measured values are more likely to be close to a value a than near the bounds of an interval with mean equal to a).

Having estimated the individual uncertainties, expressed as standard uncertainties, the next step is to combine them in the form of a *combined standard uncertainty*. Although in practice there may exist correlations between the individual uncertainty sources, however, it is usually impossible to compute those correlations accurately. For this purpose, for convenience, we rely on an assumption of independence between the individual uncertainty sources.

In particular, let $y = f\{x_1,\ldots,x_L\}$ denote an observed variable which depends on L input quantities x_l through a functional relation $f(\cdot)$. Then, the combined standard uncertainty of y, denoted by $u_c(y)$, for independent input quantities x_l, $l=1,\ldots,L$, is given by

$$u_c(y) = \sqrt{\sum_{l=1}^{L}\left(\frac{\partial f}{\partial x_l}\right)^2 u^2(x_l)} \tag{6}$$

where each $u(x_l)$ is a standard uncertainty either of type A, or of type B, while the partial derivatives $\partial f / \partial x_l$, which are called *sensitivity coefficients*, quantify how much the output y varies with changes in the values of the input quantities x_l, $l=1,\ldots,L$.

Finally, the combined standard uncertainty, which may be thought of as equivalent to one standard deviation, is transformed into an *overall expanded uncertainty*, U, via multiplication with a coverage factor k, that is,

$$U(y) = k \cdot u_c(y) \tag{7}$$

where the value of k is determined in terms of the desired confidence level, as shown in Table I.

TABLE I. COVERAGE FACTOR AS A FUNCTION OF CONFIDENCE LEVEL FOR THE GAUSSIAN DISTRIBUTION

Coverage factor (k)	Confidence level (%)
$k = 1$	67%
$k = 1.96$	95%
$k = 2.576$	99%
$k = 3$	99.7%

The most convenient way to summarize all the identified sources of uncertainty and subsequently to compute the overall uncertainty is by means of *spreadsheet tables*. An example of such a table for a temperature sensor is shown in Table II.

TABLE II. EXAMPLE OF A SPREADSHEET TABLE FOR A TEMPERATURE SENSOR

Source of uncertainty			**Value** (±)	**Probability distribution**	**Divisor**	**Standard uncertainty** $u(x)$
Type B	Sensor	Calibration	C_1	Normal	2	$C_1/2$
		Precision (Resolution)	C_2	Rectangular	$\sqrt{3}$	$C_2/\sqrt{3}$
		Sensitivity	C_3	Rectangular	$\sqrt{3}$	$C_3/\sqrt{3}$
	Sensor density		C_4	Rectangular	$\sqrt{3}$	$C_4/\sqrt{3}$
	Sampling		C_5	Rectangular	$\sqrt{3}$	$C_5/\sqrt{3}$
Type A	Temperature		C_T	-		σ_T
	Pressure		C_P	-		σ_P
Combined standard uncertainty $u_{c,b}(y)$						
Coverage factor k_b						
Expanded uncertainty U_b						

C. *Step 3: Computing correlations on uncertain data streams*

One of the major tasks of the HYDROBIONETS data management system is the correlation monitoring between uncertain data streams. Eqs. (2)-(4) are not applied directly on the raw data streams, but on the original recordings by also accounting for their estimated uncertainty. This also affects the choice of the thresholds used to decide whether two streams are "similar" or not. Specifically, the threshold $t_{new} = t_{threshold} + e$ is set based on the streams $\boldsymbol{s}_1 \pm U_1$ and $\boldsymbol{s}_2 \pm U_2$, where U_1 and U_2 are the corresponding estimated uncertainties of the two streams. Therefore, Eq. (4) is rewritten as follows:

$$p_{sim}(s,y) = \frac{1}{n}\sum_{i=1}^{n}\left[1 - \frac{\left|\tilde{S}_i - \tilde{Y}_i\right|}{2 \cdot \max\left(\left|\tilde{S}_i\right|, \left|\tilde{Y}_i\right|\right)}\right] \qquad (8)$$

where $\tilde{S}$, $\tilde{Y}$ are the truncated DFTs of the uncertain streams $\tilde{s} = s + U_s$ (or $\tilde{s} = s - U_s$) and $\tilde{y} = y + U_y$ (or $\tilde{y} = y - U_y$), respectively, with U_s and U_y denoting the uncertainties estimated in the current window of s and y, respectively.

V. EXPERIMANTAL EVALUATION

Our proposed framework is evaluated on a real dataset provided by ACCIONA Agua. In particular, the dataset consists of 22 sensors of several types (pressure, temperature, conductivity, turbidity, pH, flow, and redox), while the corresponding measurements cover a period of 1 month at a sampling rate of one measurement every three minutes. Full sensor specifications (such as, sensor precision, sensitivity, and resolution), along with the corresponding measurements were provided for each individual sensor. The inherent overall uncertainty of the recorded sensor data is quantified over sliding windows. In the subsequent results, the window size is set equal to 80 samples, which corresponds to a time interval of approximately 4 hours, while the step size is fixed at 1 sample corresponding to a time-step of about 3 minutes. The expanded uncertainty is computed by fixing the coverage factor at k = 1.96, which is equivalent to a 95% confidence level.

Fig. 2 shows the overall estimated uncertainty for a conductivity sensor. This is an example where the standard deviation is the dominant source of uncertainty. The figure also reveals an additional potential use of the estimated uncertainty as an alerter of abnormal behavior. Indeed, the time instant where the uncertainty presents a peak coincides with the time instant where the conductivity measurements deviate a lot from the previously recorded values.

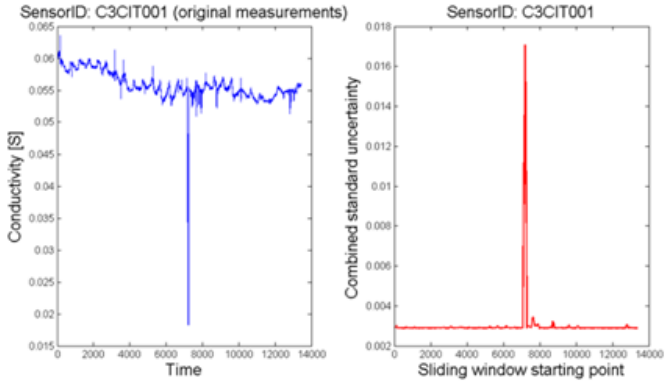

Figure 2: From left to right: a) original data; b) estimated uncertainty (window size = 80, step size = 1) for a conductivity sensor.

Concerning the performance of our proposed online correlation monitoring approach, Fig. 3 depicts the correlation and peak similarity values between a pressure sensor stream and other types of streams (the labels Prx, Tx, FFx, FLx, PHx, Cx, BFx, TRx denote pressure, temperature, feed flow, filtrate, flow, Ph, conductivity, backwash flow and turbidity sensor streams, respectively). As it can be seen in Fig 3(a) the peak similarity value is greater than or equal to the correlation value for the same pair of streams, which motivated the heuristic rule for setting the new threshold $t_{new} = t_{threshold} + e$, as mentioned in Section III.A. Fig 3(b) shows the peak similarity and the correlation values between recordings with their estimated uncertainty.

The behaviour of the error between the correlation and peak similarity values across time and over the 15 sensor pairs is shown in Fig. 4. Specifically, Fig. 4(a) shows the pairwise errors between a pressure sensor and the other 15 sensors, which is relatively low ranging between 0.02 and 0.045. In Fig. 4(b) the difference between the peak similarity value and the correlation coefficient is shown for a pair of pressure and temperature sensors, as a window of size 80 is shifted in time

for all the available measurements (13500 values). This variation in the difference between the peak similarity value and the correlation coefficient raises the need for a varying threshold t_{new}. A more efficient rule for setting this threshold adaptively is of high importance, and left as a future extension.

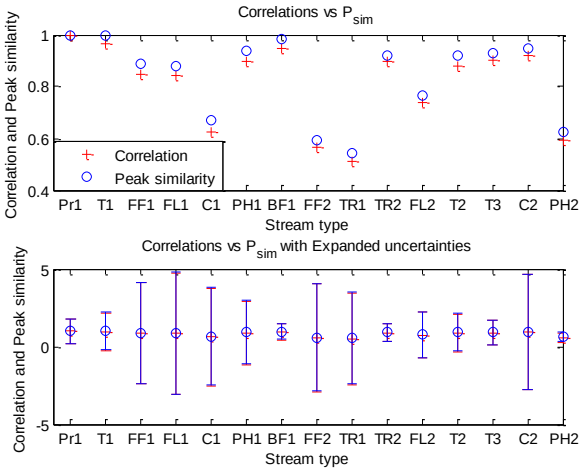

Figure 3: Correlation and peak similarity values for 15 pairs (a) of streams formed by a pressure and other types of sensors (b) with their estimated uncertainties.

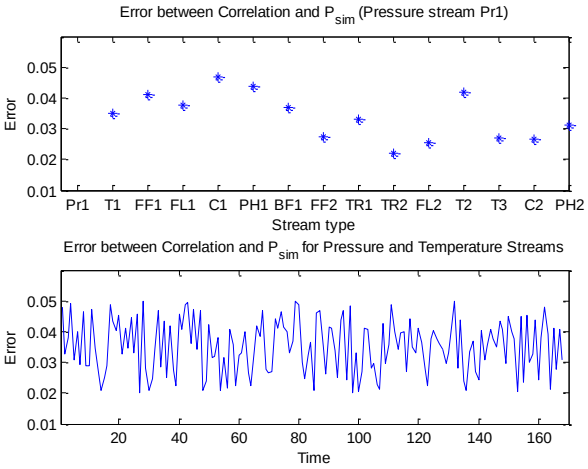

Figure 4: Error between peak similarity and correlation coefficient (a) for 15 pairs between a pressure and other different types of sensors and (b) for a single pair of streams (pressure and temperature) as the window (of size 80) is shifted in time.

Finally, in Fig. 5 we can see the significant improvement in execution time achieved by our proposed approach when compared with the typical correlation coefficient, as the length of the streams increases. Clearly, the execution time of our method remains almost constant for the selected range of length values, in contrast to the execution time for computing correlations, which increases rapidly as the stream length increases.

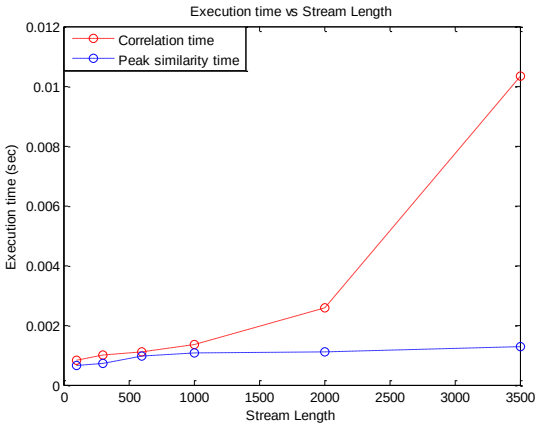

Figure 5: Execution time for peak similarity measure and correlation coefficient as a function of the stream length.

VI. CONCLUSIONS AND FURTHER EXTENSIONS

Monitoring correlations between data streams in an efficient online fashion is a significant challenge when managing data recorded in large-scale infrastructures. In this paper we proposed an approach for the design of an uncertainty-aware data management system with the ability of online monitoring correlations at a reduced computational cost. This was done by exploiting the data compression property of the DFT, in conjunction with an appropriate peak similarity measure.

As a final outcome, we envisage to provide a set of data services to manipulate sensor measurements in large-scale industrial infrastructures, as well as to identify appropriate monitoring tools for the characterization of the generated data quality in real time. As a further extension, we will focus on the design of an automatic rule for the time-varying adaptation of the threshold t_{new}, as well as the design of novel similarity measures approximating the behavior of the correlation coefficient.

REFERENCES

[1] T. Tran *et al.*, "PODS: A new model and processing algorithms for uncertain data streams," *Proc. ACM SIGMOD*, Indianapolis, June 6 – 11, 2010.

[2] T. Tran *et al.*, "CLARO: Modeling and processing uncertain data streams," *The VLDB J.*, Vol. 21, No. 5, pp. 651-676, 2012.

[3] M. Yeh, K. Wu, P. Yu, and M. Chen. PROUD: a probabilistic approach to processing similarity queries over uncertain data streams. In EDBT, pages 684–695.ACM, 2009.

[4] Y. Zhu and D. Shasha, "StatStream: Statistical monitoring of thousands of data streams in real time," *Proc. 28th VLDB*, Hong Kong, China, Aug. 20 – 23, 2002.

[5] M. Datar, A. Gionis, P. Indyk and R. Motwani, "Maintaining stream statistics over sliding windows," *Proc. SODA*, San Francisco, CA, Jan. 6 – 8, 2002.

[6] J. Gehrke, F. Korn , D. Srivastava, "On computing correlated aggregates over continual data streams," *Proc. ACM SIGMOD*, Santa Barbara, CA, May 21 – 24, 2001.

[7] J.-P. Calbimonte, H. Jeung, O. Corcho, "Querying semantically enriched sensor observations," Heraklion, Greece, May 29 – June 2, 2011.

[8] B.Canton, *Mathematics of Data Management*, McGraw-Hill ,2002

[9] C. Aggarwal, Managing and mining uncertain data, Springer, 2009.