IoT for Water Management: Towards Intelligent Anomaly Detection

Aurora González-Vidal, Jesús Cuenca-Jara and Antonio F. Skarmeta
Department of Information and Communication Engineering
University of Murcia
Murcia, Spain
{aurora.gonzalez2, jesus.cuenca1, skarmeta}@um.es

***Abstract*—Given that the global water system is deteriorating and the supply and demand are very dynamic, smart ways to improve the water management system are needed so that it becomes more efficient and to extend the services provided to the citizens leading to *smart cities*. One of many water related problems that can be addressed by the Internet of Things is anomaly detection in water consumption. The analysis of data collected by smart meters will help to personalize the feedback to customers, prevent water waste and detect alarming situations. Water consumption data can be considered as a time series. Time series anomaly detection is an old topic but in this work we attempt to examine which techniques suits better for water consumption. We examine two very well-known methods for time series anomaly detection: an ARIMA-based framework anomaly detection technique which selects as outliers those points no fitting an ARIMA process and also a technique named HOT-SAX which represents windows of data in a discrete way and then discriminates them using a heuristic. They are both very different in nature but the true positive analysis is excellent. The challenge remains in removing the false positive from the picture.**

***Index Terms*—anomaly detection, smart cities, water management, intelligent data analysis techniques**

I. Introduction

The global water system is deteriorating due to aging and stress. Furthermore, the increase in water demand is leading to infrastructure insufficiency.

Every year, more than 32 billion cubic meters of treated water from urban supply systems around the world are lost, half of which occurs in developing countries [1].

The dynamics of the supply and demand of water make it critical for governments to evaluate and better manage the water supply, requiring a smarter approach to deliver improved outcomes across the water management lifecycle. This is where the Internet of Things (IoT) takes center stage.

Thanks to the IoT, the interconnection and communications capabilities that are nowadays embedded in almost everything (places, things, people) allow to share information in a non-intrusive and efficient way. Through data analytics such information can be used to extract the knowledge that is needed to take immediate action in decisions that involve city management. When the urban infrastructure evolves using IoT, a smart city emerges. Smart cities are efficient in the way they use the resources, that includes energy [2], [3] and water.

978-1-5386-4980-0/19/$31.00 ©2019 IEEE

The deployment of a smarter approach to water management will solve some of the fundamental flaws that are provoking the worlds water crisis.

Smart water meters periodically collect measurements of water consumption. This data, stored as time series, opens up a wealth of opportunities for water providers when empowered by data mining techniques. Anomalies in water consumption time series are particularly interesting as they lead to reveal leaks, device/meter failure, detect illegal water use, warning situations and peak water use patterns in order to provide personalized feedback to customers and facilitate capacity planning and policy implementation.

II. Related Work

The state of the art academic research includes the use of machine learning algorithms in tasks related to all the steps in the urban water process: from the configuration of the architecture of the network to managing the network using prediction and anomaly detection.

The division of large water systems enables better water management by improving efficiency and safety through strategic rule implementations. For such task using manual or empirical approaches is outdated since the several physical and hydraulic features that are available nowadays thanks to the IoT can serve as precise attributes. In [4] graph clustering is proposed in order to optimally create the District Metered Areas (DMAs). The status of the boundary valves and their location at DMA entrances is also optimized in the mentioned work. Their process is based on three algorithms: Genetic Algorithms (GAs), Particle Swarm Optimisation (PSO), and Soccer League Competition (SLC).

Having defined the network according to data, further analysis of the information that is gathered by smart meters can be done in order to forecast water demand. Water demand forecasting is of utmost importance for the management of Water Distribution Systems (WDSs) due to its close relationship with the operational and economic aspects of water distribution, and also because this demand introduces high levels of uncertainty in WDS hydraulic models. When developing accurate forecasting models from smart meter readings, a preprocessing study that includes feature selection or data transformation must be done. Principal Component Analysis (PCA), Self-Organizing Maps (SOMs) and Random Forest (RF) algorithms ares tested in [5] for exploring weather, social and economical variables. Those feature selection techniques are used in order to select the proper variables to build hourly regression models. Also, Discrete Wavelet Transform (DWT) is applied to transform the data in [6] in order to carry out a monthly consumption prediction using nonlinear techniques such as Artifical Neural Networks (ANN).

As water distribution networks present wear and tear, efficient real-time assessment and monitoring is crucial for an optimal functioning. Damages are most often manifested as pipe-failure incidents and lead to significant levels of non-revenue water (typically in the range 20%-30%) [7] . Traditional approaches average periodic consumption and compare this to the one from a corresponding past period in order to find anomalies. This methodology does not adapt to changes in the time series, does not consider drifts and it can not be used in real time. The real-time operability of the methods for anomaly detection is important since the decision-making process is sometimes carried on in a real-time way. Given that water consumption measurements per metering unit can be considered as times series, change-point methods are suitable for the detection of anomalies. In [7], using such methods allow to distinguish the kind of anomaly that is being detected: a discontinuity in the signal (a break in the consumer's water consumption patterns) or an unusual increase in the signal (waterloss incidents). Probabilistic outlier detection has also been explored [8] in water management scenarios. It requires a probability distribution for data, in which data points assigned a low probability are judged as outliers. This probability distribution can be represented by a Deep Neural Network (DNN) that is trained on normal data. In [9] an approach based on using GAs to find the best neuronal network architecture for a given dataset is introduced. In order to reduce the false-negative ratio, they also use exponentially weighted moving average smoothing, mean p-powered error measure, individual error weight for each tag values and disjoint prediction windows.

Heatwave events in temperature time series have been studied in [10] and since they can also be considered anomalies, we find their approach very interesting. The authors propose a multiresolution quantile (MRQ) approach, extending a variation of the common SAX methodology (that is further explained in Section III) and computing the quantiles for each level of resolution in the time series. MRQ starts dividing the series into segments of equal length using piecewise aggregate approximation (PAA). Then, SAX runs on the quantiles of each segment of the data. Finally, the differences between upper and lower quantiles at each resolution level is computed by a lower-bounding distance measure. Anomalies (heatwave events) could then be detected based on the persistence of minimum distances at various time-based resolution levels.

III. Methodology

Anomaly detection for time series has been widely studied in the past. It is our duty to discover the weak and strong points of methods from different nature to find anomalies specifically in water consumption time series in order to test their goodness.

For this task, we propose a two-steps scheme that firstly will extract outliers and abnormal patterns using the individual time series properties of the data and secondly, using the features extracted by such models will try to classify them thanks to the annotated classes. The second step adds extra value to the process since uses knowledge in order to discard false positives, which are very common in this area.

A. Step 1: Individual time series anomaly detection

We have tested two different approaches: an ARIMA based algorithm and a heuristic that uses Symbolic Aggregate Approximation (SAX) for finding time series discords.

- ARIMA
  In order to locate the time series outliers, the ARIMA based framework described in [11] is used. Here, five types of outliers are considered: "AO" additive outliers, "LS" level shifts, "TC" temporary changes, "IO" innovative outliers and "SLS" seasonal level shifts.
  Let z_t be an ARMA model, then we assume that our series y_t, which contains m outliers can be described as:
  $$y_t = z_t + \sum_{j=1}^{m} w_j L_j(B) I_{t_j},$$
  where $I(t_j)$ is an indicator variable with value 1 at the time t_j, that is when the outlier $L_j(B)$, with weight w_j arouses.
  For further information about how each of the outliers is defined, please check the implementation manual [12] or the original paper where this methodology was proposed [11].
- Heuristically Order Time series using SAX (HOT-SAX)
  The large volume of data that is collected by means of IoT can be aggregated and represented in efficient and higher-granularity ways. The idea is to create sequences of patterns and data segments that occur in large-scale IoT data streams. In order to reduce the number of data points in a series and create a representation, segmentation and representation methods are advised [13].
  Among all the techniques that have been used to reduce the number of points of a time series data, Symbolic Aggregate Approximation (SAX) has specially attracted the attention of the researchers in the field.
  Using SAX, a time series is normalised and then discretized by first obtaining a Piecewise Aggregate Approximation (PAA), that is dividing the original data into the desired number of windows and calculating the average of data falling into each window. Secondly, predetermined breakpoints are used to map the PAA coefficients into

symbols or letters creating words. That is, each of the windows is a word.

In that sense, it is possible to use a heuristic search for finding the words that appear to be less frequent in our time series and define such segments as anomalies [14].

B. Step 2: Anomaly and discord classification

Anomalies obtained using each of the methods are different: ARIMA provides specific points (see red points of column 1 in Fig. 2) and the outlier type and HOT-SAX provides a series of points that construct a subseries (see red subseries of column 2 in Fig. 2).

At this point, we want to use algorithm that determine if some combination of points or specific subseries are indicators of the anomalies that so far have been annotated.

- Association Rule Learning for ARIMA framework
  With the aim of trying to differentiate which types of outliers correspond to anomalous behaviors that produce breakdowns and which correspond to normal fluctuations in consumption, we have tested the ARIMA algorithm in the set of consumption series, differentiating two classes: those in which breakdowns are contained and those in which they are not (we consider that there is a period without breakdowns as long as there is not a breakdown in a timeframe of at least two months with respect to the period analyzed). As a result of applying the ARIMA agloritm, a set of outliers of different types (described previously in section 3.1) were obtained for periods with breakdowns and without breakdowns. An Association Rule Learning Algorithm was applied to these sets of outliers with the aim of studying if the type of outliers detected by the algorithm is representative for each class. In order to carry out a more exhaustive study, the Association Rule Learning Algorithm was applied on different size sets: taking only the outlier closest to the breakdown, the two closest, the three closest etc.
- Subseries classification using Random Forest for HOT-SAX discordances
  As we set the parameters of HOT-SAX algorithm to find n discordances per series we decided to separate series that lead to anomalies and series which does not lead to anomalies and apply the algorithm. In that case, we obtained n subseries per series that can be classified as 0 (belonging to normal) or as 1 (leading to anomaly).
  After that, we apply a random forest classifier [15] in order to fit a model that discriminates between subseries that indicate a future anomaly and subseries that do not.

C. Metrics

Accuracy is simply the ratio of number of correct predictions to the total number of input samples, without making any difference between correctly classified anomalies and correctly classified not anomalies.

Due to the nature of our problem, we are going to use diagnostic accuracy metrics in order to measure the agreement between the predicted class (abnormal or normal) and the annotated anomalies that have occurred.

We have decided to go further than just reporting accuracy since for the special case under study it is much more essential to find the real abnormal situation than to claim a abnormality when actually is not.

The 2 x 2 contingency table gathers the index test results on one side and those of the reference standard on the other.

TABLE I
CONTINGENCY TABLE DEFINITIONS

	Anomaly	Not anomaly
Anomaly predicted	True positive (TP)	False positive (FP)
Not Anomaly predicted	False negative (FN)	True negative (TN)

Sensitivity ('positivity in anomaly") is the proportion of subseries which belong to the abnormal one and give positive test results. Shortly, right classified abnormal subseries.

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity ('negative in anomaly") refers to the proportion of subseries that belong to the normal one and give negative test results. Shortly, right classified normal subseries.

$$Specificity = \frac{TN}{TN + FP}$$

The accuray of a test to discriminate abnormal from normal cases is sometimes evaluated using Receiver Operating Characteristic (ROC) curve analysis. The area under the ROC curve (AUC) determines the ability of the model to discriminate between the normal and abnormal subseries.

For our problem is specially important to obtain a high sensitivity. This means that abnormal data should not be ignored. Even if some false positives are encountered it is way more important to detect all abnormal situations.

IV. USE CASE AND EXPERIMENT RESULTS

Located in Spain, region of Murcia has designed and deployed a water management platform to monitor and take decision over urban water distribution systems integrated in a smart city approach. The pilot combines different sources of information including smart metering system, company's SCADA with real time information about pressures (see Fig.1), water quality, and GIS system providing infrastructure deployment and maintenance historical records, as data from external sources (e.g. weather, financial conditions, seismic activity).

Over this baseline several big data analytics components will be deployed that will share common data models and interoperability capabilities.

The analytic task that is presented in this work focuses on anomaly detection, having a wide range of applications as fraud detection, surveillance, diagnosis, data cleanup, and predictive maintenance.

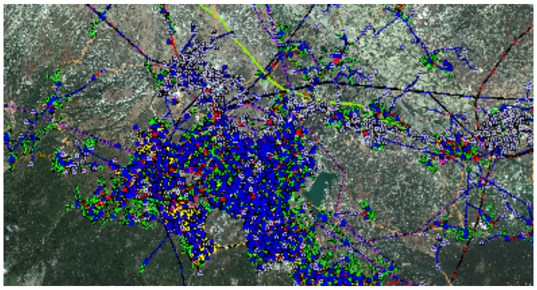

Fig. 1. SCADA real time pipe information

The platform gathers data from 40000 smart meters that mainly belong to factories and buildings (owners' communities). Every data point is gathered with a timestamp, that is, it has the form of a time series.

A. Experiments

We are given a very reduced annotated dataset of anomalies. This task is developed nowadays by a human and it is done exclusively on the users that consume the most, which are some factories. Since the goal of this study is to create a baseline in order to develop a tool able to warn an expert about possible abnormal situations, we consider that finding an outlier close to the real anomaly is a success. This is because at this point, the expert will be able to validate the finding and take further actions.

That way, we have tested the algorithms on the 30 time series which were anotated. Originally, consumption is an accumulated value and the measurements are irregular. We have aggregated consumption each 2 hours and we have considered 1 month measurements in order to carry out the analysis. In average and after the aggregation phase, 400 observations per water meter are used in order to detect the anomaly.

The first step results of analysing 3 water meters are shown in Fig. 2. In such figure, rows represent the water meter and the 2 columns are showing the results using both previously stated methods. In all 6 graphs, the vertical red line represents when the human expert was able to find the anomaly. In the graphs relative to the ARIMA framework outliers are represented with red dots, and while the original series is grey, blue is the corrected one (the values that are selected as outliers are corrected, but this does not interest us). In the graphs relative to HOT-SAX, the original series is blue and the most abnormal data sequences are highlighted in red.

Both tested approaches seem to be good at detecting the anomalies: 90 % are found using the so called ARIMA-framework, and 80 % using HOT-SAX. However, both approaches present a high rate of false positives, that is, they find anomalies which have not been stated as so by the expert. This way, the second step of our methodology is not only justified but also necessary.

For analysing HOT-SAX results, we have stated to find 5 discord subseries of length 20 per time series. Using 70 % of the subseries as training and 30 % as test we have been able to obtain a 76 % accuracy when classifying them. More important than that, sensitivity is 86 % and specificity is 72 %. All this values can be calculated from Table II.

TABLE II
CONTINGENCY TABLE RESULTS

	Anomaly	Not anomaly
Anomaly predicted	12	9
Not Anomaly predicted	2	23

Finally, the results obtained using the Association Rule Learning are inconclusive, mainly due to the lack of data because the more outliers we take near the breakdown the fewer examples we obtain. This fact is given for obvious reasons because differently to HOT SAX algorithm, that always finds discordances, in ARIMA algorithm outliers are not always detected and even in case of detecting them the number of outliers obtained varies considerably. This inversely proportional relationship makes it very difficult to get consistent results, and consistently the results obtained, in addition to showing few significant differences between the series with and without breakdowns, lack reliability. However for future work the infrastructure deployed for this study can be very useful in databases with a higher number of data, in which is possible to extract relevant information that can clarify wich factors determine the causes of breakdowns .

V. Conclusions and Future Work

In this preliminary study we have tested how two very different algorithms in nature can be used to discover anomalies on water consumption time series. To our knowledge, this is the first time that a combination between association rules and time series anomaly detection is used for discarding false anomalies, which are very common in this problem. We consider that our work is a step forward towards automatic water management in smart cities. In many utility companies, anomaly detection is either neglected or done by a technician who normally is unable to check all smart meters due to the huge amount of consumers that are connected to the network and the high volume of data that is generated. This is also the case of our particular pilot, where a person does this work manually having to reduce the search dataset to a subset of very high consumers and also being unable to anticipate or to detect the anomaly in real time.

Since the dataset used for testing the models is obtained as above described, further work needs to be done. In that sense we propose the following:

- Select other algorithms based on a machine learning approach instead of focusing just on time series and compare results.
- Design an experiment which tests several anomaly detection models and finds the best one through expert validation. That is, the expert validates the findings of the algorithms and the outcomes are stored in order to

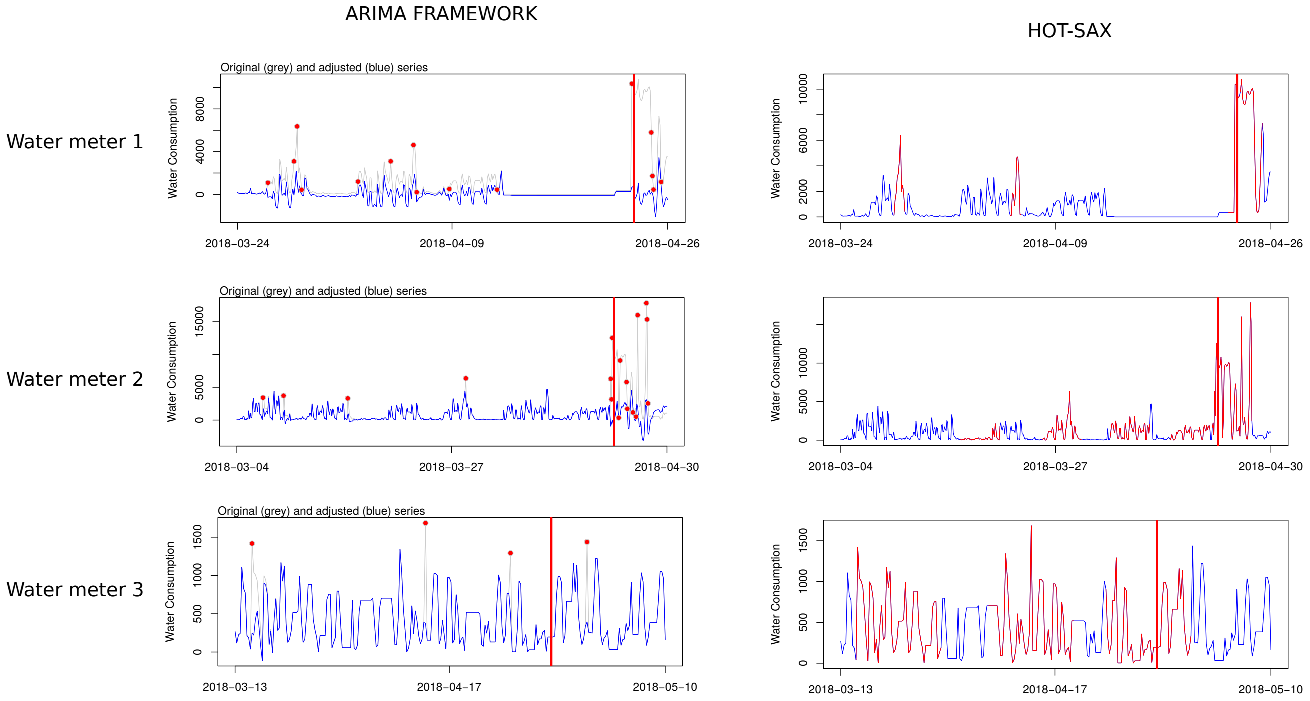

Fig. 2. Anomaly detection results on 3 smart meters

select the best one for the pilot case in particular and for the problem in general. That can be done by incorporating other publicly avaiable water consumption datasets.

- Elaborate a friendly graphic user interface that can be understandable and usable by the technician in order to validate the previous.
- Analyze and develop retrofit techniques which can be used in order to improve or change dynamically the model parameters (retrain) using the expert input in order to achieve a fully automatic procedure.
- Use Big Data frameworks in order to facilitate the analysis on real time.

Acknowledgment

This work has been sponsored by MINECO through the PERSEIDES project (ref. TIN2017-86885-R) and grant BES-2015-071956 and by the European Comission through the H2020-ENTROPY-649849 and the H2020 IoTCrawler (contract 779852) EU Projects.

References

[1] B. Kingdom, R. Liemberger, P. Marin, The challenge of reducing non-revenue water (NRW) in developing countries. How the private sector can help: A look at performance-based service contracting, World Bank Group, 2006.

[2] M. V. Moreno, F. Terroso-Sáenz, A. González-Vidal, M. Valdés-Vela, A. F. Skarmeta, M. A. Zamora, V. Chang, Applicability of big data techniques to smart cities deployments, IEEE Transactions on Industrial Informatics 13 (2) (2017) 800–809.

[3] A. González-Vidal, A. P. Ramallo-González, F. Terroso-Sáenz, A. Skarmeta, Data driven modeling for energy consumption prediction in smart buildings, in: Big Data (Big Data), 2017 IEEE International Conference on, IEEE, 2017, pp. 4562–4569.

[4] B. Brentan, E. Campbell, T. Goulart, D. Manzi, G. Meirelles, M. Herrera, J. Izquierdo, E. Luvizotto Jr, Social network community detection and hybrid optimization for dividing water supply into district metered areas, Journal of Water Resources Planning and Management 144 (5) (2018) 04018020.

[5] B. M. Brentan, G. Meirelles, M. Herrera, E. Luvizotto, J. Izquierdo, Correlation analysis of water demand and predictive variables for short-term forecasting models, Mathematical Problems in Engineering 2017.

[6] A. Altunkaynak, T. A. Nigussie, Monthly water consumption prediction using season algorithm and wavelet transform–based models, Journal of Water Resources Planning and Management 143 (6) (2017) 04017011.

[7] S. Christodoulou, E. Kourti, A. Agathokleous, Waterloss detection in streaming water flow timeseries using change-point anomaly methods.

[8] J. Inoue, Y. Yamagata, Y. Chen, C. M. Poskitt, J. Sun, Anomaly detection for a water treatment system using unsupervised machine learning, in: Data Mining Workshops (ICDMW), 2017 IEEE International Conference on, IEEE, 2017, pp. 1058–1065.

[9] D. Shalyga, P. Filonov, A. Lavrentyev, Anomaly detection for water treatment system based on neural network with automatic architecture optimization, arXiv preprint arXiv:1807.07282.

[10] M. Herrera, A. A. Ferreira, D. A. Coley, R. R. de Aquino, Sax-quantile based multiresolution approach for finding heatwave events in summer temperature time series, AI Communications 29 (6) (2016) 725–732.

[11] C. Chen, L.-M. Liu, Joint estimation of model parameters and outlier effects in time series, Journal of the American Statistical Association 88 (421) (1993) 284–297.

[12] J. L. de Lacalle, tsoutliers: Detection of Outliers in Time Series, r package version 0.6-6 (2017).
URL https://CRAN.R-project.org/package=tsoutliers

[13] A. Gonzalez-Vidal, P. Barnaghi, A. F. Skarmeta, Beats: Blocks of eigenvalues algorithm for time series segmentation, IEEE Transactions

on Knowledge and Data Engineering.

[14] E. Keogh, J. Lin, A. Fu, Hot sax: Efficiently finding the most unusual time series subsequence, in: null, Ieee, 2005, pp. 226–233.

[15] H. Deng, G. Runger, E. Tuv, M. Vladimir, A time series forest for classification and feature extraction, Information Sciences 239 (2013) 142–153.