Smart water in urban distribution networks: limited financial capacity and Big Data analytics

A. Candelieri[1] & F. Archetti[1,2]
[1]*Consorzio Milano Ricerche, Italy*
[2]*Department of Computer Science, Systems and Communication, University of Milano Bicocca, Italy*

Abstract

Big Data opportunities arise from high rate data streams acquired through smart sensors and smart meters, which, even for small water utilities, may produce a huge amount of data to be stored. This data enables the application of new data analytics to infer reliable predictive functionalities, with implications ranging from reducing No Revenue Water (NRW) to optimizing the water-energy nexus, meeting ever more pressing budgetary constraints. This paper presents the approach proposed in the EU-FP7-ICT project ICeWater, combining time series clustering, for the identification of typical daily urban water demand patterns, and Support Vector Regression for performing a short term forecast. Promising results obtained on the Water Distribution Network (WDN) in Milan are presented. The approach has been designed to also be applied on smart metering data related to individual customers, addressing Big Data analytics issues.
Keywords: smart water management, predictive analytics, short-term demand forecasting.

1 Introduction

Growing demand, aged infrastructures almost at the end of their remaining useful life, climate change, decreasing fresh water sources and quality, increasingly stringent regulations and budget constraints are requiring a pressing need for a more sustainable and efficient management of urban Water Distribution Networks (WDN). The key to achieving this improvement is "understanding *where*, *when* and *why* we use water" [1]. The adoption of

WIT Transactions on The Built Environment, Vol 139, © 2014 WIT Press
www.witpress.com, ISSN 1743-3509 (on-line)
doi:10.2495/UW140061

systems able to integrate robust and proven ICT solutions and innovative data analytics approaches for the efficient management of the water-energy nexus, such as the ICeWater project (co-funded by the European Commission). These systems enable the shift from the data-poor, hardware-centric, asset-driven XIX century business model to a data-rich, information- and customer-centric environment, supported by smart-sensors, smart-metering and on-line monitoring systems.

From a technological point of view, ICT based solutions, such as Supervisory Control And Data Acquisition (SCADA) systems are already widely adopted by water utilities in order to monitor and control the hydraulic behavior of the WDN. These technological systems are also able to generate warnings and alarms according to specified rules and to store data which can be analyzed through advanced analytics for enabling a more effective and efficient leakage management [2–4].

With respect to individual customers, Automatic Metering Readers (AMR) are quite innovative and promising ICT solutions which are gaining recent interest in the field of "smart water". However, AMRs are quite expensive with respect to SCADA as their installation involves all the customers of the WDN, rather than a few relevant monitoring points. However, the availability of a huge amount of high-rate data, related to the consumption of customers, will be the best and innovative advantage for performing more accurate customer-segmentation, "targeted" demand management strategies and individual demand forecasting.

Big Data opportunities arise from high rate data streams acquired through these smart solutions producing, even for small water utilities, a huge amount of data which may be analyzed to reduce No Revenue Water (NRW) as well as optimizing the water-energy nexus, addressing the pressing issue of the limited financial capability at WDNs.

Although the developed world has been forged on the supply-side, the historical period requires looking to curbing demand as an active, rather than reactive, water management strategy [5, 6]. With respect to this, the capability to reliably forecast demand is crucial for maintaining a satisfactory level of the service while reducing costs for caption, treatment, storage and distribution. Moreover, a demand forecast may improve WDN management at very different levels according to the time window considered: *planning*, *strategic* and *operation* levels with respect to long, medium and short term forecast.

The main contribution of this paper is related to the design and development of two specific decision support functionalities of the ICeWater project's Decision Support System, namely:

- the identification of typical urban water demand patterns
- the related forecast in the short term (today or tomorrow).

The approach has been developed and validated on historical urban water demand data retrieved from the SCADA of Metropolitana Milanese in Milan, the Italian pilot of the ICeWater project. The approach is aimed at analyzing a huge amount of time-series (water consumption over time, at hourly, and even lower,

level) and then cluster them to identify typical consumption patterns and derive reliable predictive models.

The approach has been designed to be also applied "as is" on AMR data related to individual customers, thus showing characteristics of a Big Data Analytics application. With respect to this, preliminary work is ongoing, following the installation of AMRs in the pilot site.

As result, a reliable short term demand forecasting model has been obtained for urban water demand in Milan, enabling the optimization of caption, treatment, storage and distribution by using energy (in particular for pumping) when it is less expensive during the day. A recent work [7] reports that forecasts led to a 3.1% reduction of energy consumption and a 5.2% reduction of energy costs at a WDN in The Netherlands.

The proposed *pattern-discovery-based* approach provides a reliable prediction depending on the hourly urban water demand acquired by SCADA in the first hours of the day and does not require any "on-line updating" and is not affected by the "time-lag" effect, usually occurring in more classical approaches (e.g., ARIMA).

The rest of the paper is organized as follows: section 2 describes the available data and the methodologies applied; in section 3 the proposed approach is illustrated; section 4 reports the obtained results. Some conclusions are finally provided.

2 Material and methods

The data considered in this study has been retrieved from the SCADA system of the WDN in Milan which is managed by Metropolitana Milanese (MM). MM is one of the two case studies of the EU-FP7-ICT project ICeWater; this highly interconnected WDN is shown in Figure 1.

Historical urban water demand data was retrieved for the period 01 March 2011 to 31 March 2012. Data was organized as a time series dataset, where each entry into the dataset consists of 24 measurements, which are the hourly volume of water delivered by MM over the day.

It is important to note that MM is characterized by a really low leakage level. This avoids frequent and significant distortions into the daily urban water demand time series data making more reliable the identification of typical daily consumption patterns.

As a first step, preliminary preprocessing on the retrieved data has been performed, aimed at identifying anomalous values and replacing missing values. Nonetheless, this procedure affected only a very limited portion of data due to the reliability of the SCADA system.

2.1 Time series clustering for pattern identification

A specific survey on clustering of time series data has been proposed in [8], where the basics of time series clustering are presented, including general-purpose clustering algorithms which are commonly used, criteria for evaluating

Figure 1: The WDN in Milan, Italy, managed by Metropolitana Milanese.

performance, and similarity/dissimilarity measures used to compare two time series. These considerations are general and affect even more recent studies related to the clustering of time series data stream [9, 10].

Furthermore, three different strategies are possible, by working:

- Directly with the raw data (usually in time domain, but even in frequency domain);
- Indirectly with features extracted by the raw data;
- Indirectly with models built from the raw data.

The raw-data-based strategy is different from clustering of static data in replacing the distance/similarity measure with an appropriate one for time series.

The feature-based strategy converts a raw time series data either into a feature vector of lower dimension and then applies a conventional clustering algorithm to the extracted feature vectors.

The model-based strategy is similar to the feature-based one, converting raw time series data into a number of model parameters to consequently apply a conventional clustering algorithm to these parameters.

In a most recent and interesting work about a novel clustering method on time series data [11], a more accurate distinction between the different types of similarity that could be evaluated among time series is proposed:

- Type 1: similarity in time. The goal is to cluster together series that vary in a similar way on each time step.

- Type 2: similarity in shape. The goal is to cluster together time series having common shape features.
- Type 3: similarity in change. The goal is to cluster together time series that vary similarly from time step to time step.

The identification of consumption patterns proposed in ICeWater is to provide managers with a reliable analytical tool which does not require any specific skills or competences on data analysis. The approach has been designed to address the WDN needs for a more accurate customer profiling, the identification of typical and periodic behaviors, the continuous monitoring of water consumption patterns and variations along time and customers (space).

2.2 Support Vector Machines for regression

Support Vector Machines (SVM) [12] is a well known machine learning strategy to (semi-)automatically discover, from an available set of data, a general relationship between the values of some variables of interest (features) and one target variable, by minimizing the prediction error. The regression model learned via SVM is expressed as a function of a subset of data (namely, support vectors).

SVMs had a sound orientation towards real-world applications; initial work focused on OCR (optical character recognition) and in a short period of time, SV classifiers became competitive with the best available systems for both OCR and object recognition tasks. A comprehensive tutorial on SV classifiers was published in [13]. But also in regression and time series prediction applications, excellent performances were soon obtained [14] containing a more in-depth overview of SVM regression. Additionally, [15] and [16] provide further details on kernels in the context of classification.

3 The proposed approach

The proposed approach is "completely data-driven": the idea is that variability in urban water demand, due to different consumption behaviors in seasons, days of the week, and hours of the days, is all hidden into the data and that can be extracted and characterized through machine learning. As already mentioned, the approach consists of two consecutive phases:

- the former is devoted to clustering together daily demand patterns, represented by the volume of water delivered at each hour, in order to identify the most typical patterns in consumption;
- the latter aims at identifying a prediction model, based on the Support Vector Regression, able to predict, at one time, the urban water demand at each (remaining) hour of the day, given the hourly consumption very early in the morning as acquired through SCADA.

Clustering techniques capturing similarity in shape (i.e., by using triangle similarity) and considering only the raw time series data without any other information (e.g., day of the week, season or weather data) are used for identifying typical daily consumption patterns. All the time series to analyze are

defined in the same time window (i.e., a day) and thus have the equal length (i.e., 24 data points in the case of hourly consumption data).

As a result of this step, a limited set of typical daily urban water demand patterns is identified where a "stereotype" (e.g., the mean daily urban water pattern) is defined for each cluster.

At the end of this step, a possible relationship between each stereotype/cluster and the time of its occurrence (e.g., period of the year and/or type of day) is considered in order to provide some "semantics" about the water usage behavior associated to each stereotype. It is suitable to take into account at least one year in order to capture possible seasonality. Semantics is retrieved by visualizing the distribution of the identified patterns over the days, within the analysed period, in order to evaluate possible seasonality, surprising periods, and daily/weekly habits.

Successively, each cluster is considered as a dataset and is used to learn a SVM regression model able to predict the urban water demand at a specific hour depending on the first *m* hourly data acquired through SCADA. As result, a pool of SVM regression model is generated for each cluster.

This procedure is summarized, with respect to a specific cluster, in Figure 2.

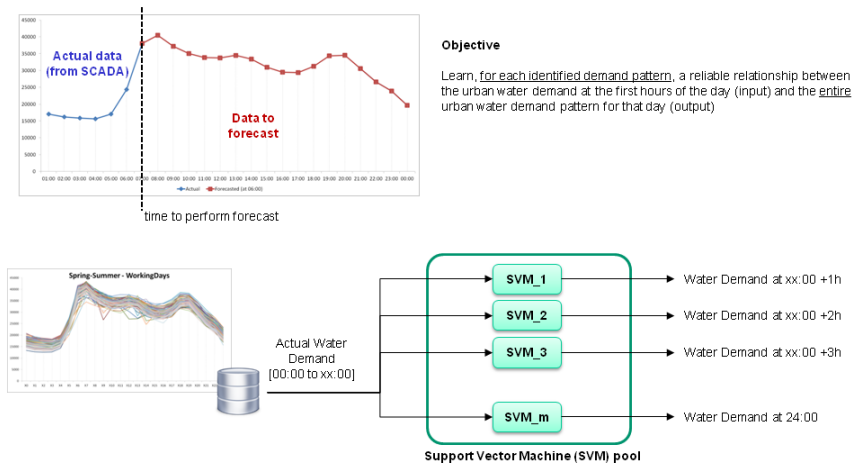

Figure 2: Learning predictive models: one pool of SVM regression model for each typical pattern identified; one SVM regression model for each hour.

The pools of SVM regression models are stored; the most suitable pool is identified and retrieved, then the correspondent models are used to predict the hourly water demand data given the first *m* values acquired through SCADA.

Figure 3 shows this procedure by using only the first 6 hourly values as input of all the models in the selected pool.

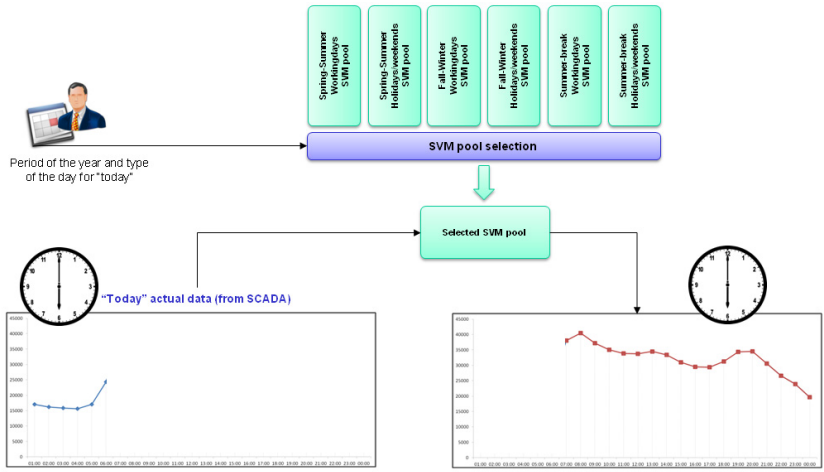

Figure 3: Applying predictive models learned: the most suitable pool of SVM regression models is selected (i.e., depending on the period of the year and type of the day) and each model is used to forecast urban water demand at each hour.

4 Results

In this section the main results obtained are presented. Figure 4 shows the 6 typical daily urban water demand patterns (stereotypes) identified on the data from MM's SCADA system, computed as the average of time series in the correspondent cluster.

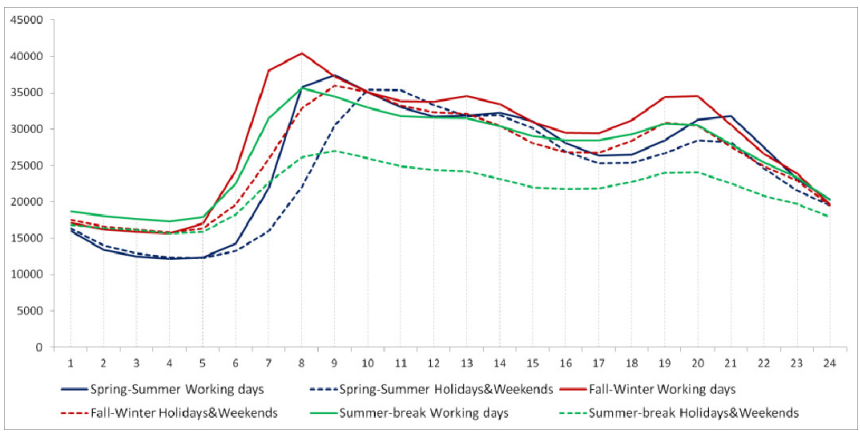

Figure 4: Typical patterns identified in the urban water demand data of the WDN in Milan.

As every time series in a cluster is related to a specific day, it has been possible to make some considerations about each cluster (and correspondent stereotype) by looking at the distribution of the clusters over the analysed time period.

In particular, three different periods of the year have been identified (namely, Spring-Summer, Fall-Winter and Summer-break) and 2 different types of day for each time period (namely, working-days and holiday-weekends).

It is really easy to note that major differences regarding the peaks in consumption in the morning, as well in the evening, both for period of the year and type of day. In particular, the peak in the morning is always delayed by about 1 hour for each period of the year. Moreover, the stereotypes named "Summer-break – working-days" is a really specific daily urban water demand pattern, more "flat" and "low" and it is associated to the 15 days in the middle of August, when usually citizens have their summer holidays and leave Milan.

The identified clusters have been then used for training the SVM regression models by using the first 6 values of hourly consumption as input features. One SVM has been trained for each hour of the day (from the 7th to the 24th), that is the target variable, and for each cluster. Forecasting performances have been evaluated through leave-one-out validation, in order to estimate the reliability of the predictions on new coming time series data. Several possible configurations for each SVM regression model have been taken into account, using both Polynomial and Radial Basis Function (RBF) kernels.

As a reliability index, the (absolute) percentage error has been computed in correspondence with each hour (i.e., $| \textit{actual} - \textit{predicted} | / \textit{actual}$). This value is then averaged on all the hours to predict and the result has been used to select the most reliable SVM configurations.

Table 1 reports the average error and its standard deviation for the best and the worst forecasts, on each cluster. Finally, the following figures show the best and the worst forecasts for each cluster.

Table 1: (Absolute) percentage error for the best and the worst forecasts in each cluster; mean and standard deviation of the error over the day.

	Best		Worst	
	Mean	StdDev	Mean	StdDev
Cluster 1	0.79%	0.59%	6.11%	2.95%
Cluster 2	1.57%	1.18%	14.33%	11.68%
Cluster 3	0.84%	0.66%	8.48%	3.53%
Cluster 4	1.71%	2.56%	12.84%	7.53%
Cluster 5	1.31%	0.93%	7.85%	13.26%
Cluster 6	1.10%	0.85%	6.54%	3.46%

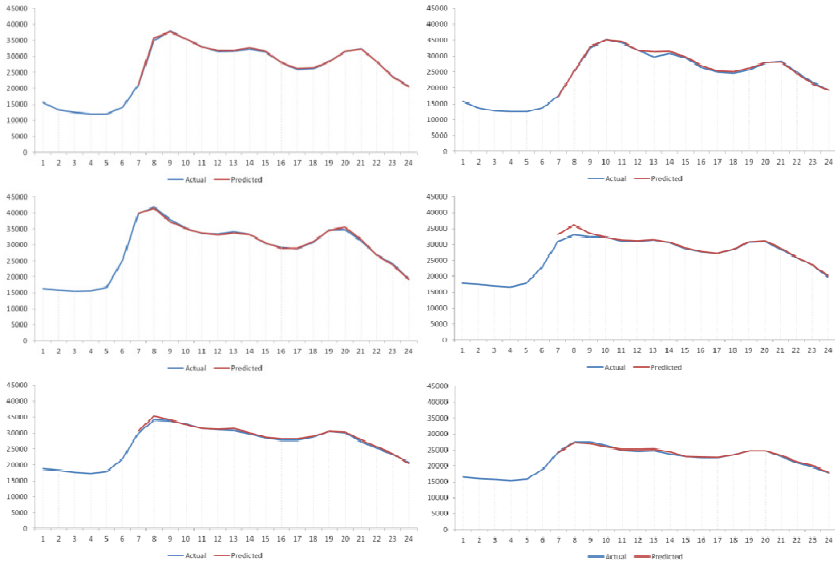

Figure 5: Best forecast for each cluster (leave-one-out validation).

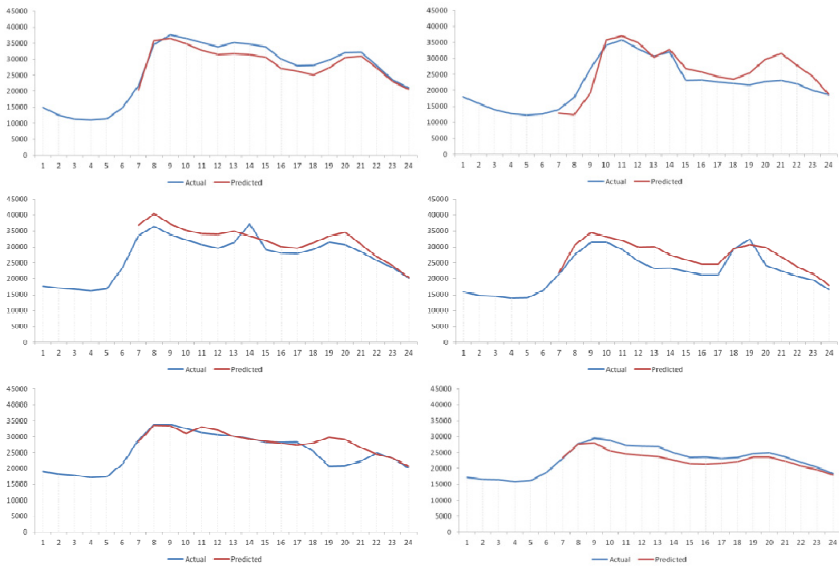

Figure 6: Worst forecasts for each cluster (leave-one-out validation).

5 Conclusions

The approach presented in this paper, developed within the EU-FP7-ICT project ICeWater, proposes a predictive analytics solution for short term water demand forecast. The promising results obtained by its application on real data retrieved from the SCADA system of Metropolitana Milanese, the WDN in Milan and one of the two use cases of ICeWater, proved that the proposed combination of time series data clustering and Support Vector Machine regression is effective in implementing a completely data-driven approach to identifying typical consumption patterns and performing reliable predictions in the very short term (today or tomorrow).

While typical consumption patterns' identification, in particular when applied at an individual customers level, enables a better segmentation of the users and supports the definition of demand management strategies for improving water and costs savings per se, it also permits a highly reliable forecast of the water demand which can be used to effectively optimize pumping scheduling and storage, reducing energy costs for caption, treatment, storage and distribution.

The approach has been designed to be applicable even on AMRs data; a study about that is currently ongoing. The availability of these ICT solutions will largely intensify the requirement of Big Data solutions, for data management as well as analytics. The proposed approach has been developed to be scalable and *runnable* on parallel/distributed architectures, qualifying itself as a Big Data Analytics approach for supporting smart water in modern cities.

Acknowledgement

This work has been partially supported by the European Union ICeWater project – FP7-ICT 317624 (www.icewater-project.eu).

References

[1] Gleick, P. H., Roadmap for Sustainable Water Resources in Southwestern North America, PNAS 107, 50: 21300–21305, 2010.

[2] Candelieri, A., Conti, D., Archetti, F., A graph based analysis of leak localization in urban water networks, 12th International Conference on Computing and Control for the Water Industry, CCWI2013, 2013a.

[3] Candelieri, A., Archetti, F., Messina, E., Improving leakage management in urban water distribution networks through data analytics and hydraulic simulation. WIT Transactions on Ecology and the Environment 171, 107–117, 2013b.

[4] Candelieri, A., Messina, E., Sectorization and analytical leaks localizations in the H_2O Leak project: Clustering-based services for supporting water distribution networks management. Environmental Engineering and Management Journal 11(5), 953–962, 2012.

[5] Hill, T., Symmonds, G., The Smart Grid for Water: How Data Will Save Our Water and Your Utility, Ingram Pub Services 2013.

[6] Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., Stouffer, R. J., Stationarity is dead: whither water management?, Science 319, 2008.
[7] Bakker, M., Vreeburg, J. H. G., Palmen, L. J., Sperber, V., Bakker, G., Rietveld, L. C., Better water quality and higher energy efficiency by using model predictive flow control at water supply systems, Journal of Water Supply: Research and technology – AQUA 58 (3), 203–211, 2013.
[8] Liao, T.W., Clustering of time series data – a survey, Pattern Recognition 38, 1857–1874, 2005.
[9] Pereira, C. M. M., de Mello, R. F., TS-stream: clustering time series on data streams, Journal of Intelligent Information Systems, 2013.
[10] Kavitha, V., Punithavalli, M., Clustering Time Series Data Stream – A Literature Survey, (IJCSIS) International Journal of Computer Science and Information Security, 8, 1, 2010.
[11] Zhang, X., Liu, J., Du, Y., Lv, T., A novel clustering method on time series data, Expert Systems with Applications, 38, 11981–11900, 2011.
[12] Vapnik, V., Statistical Learning Theory, New York, Wiley, 1998.
[13] Burges, C. J. C., A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2): 121–167, 1998.
[14] Scholkopf, B., Smola, A. J., Learning with Kernels, MIT Press, 2002.
[15] Cristianini, N., Shawe-Taylor, J., An Introduction to Support Vector Machines. Cambridge University Press, Cambridge, UK, 2000.
[16] Herbrich, R., Learning Kernel Classifiers: Theory and Algorithms, MIT Press, 2002.