2014

Urban Water Demand Characterization And Short-Term Forecasting – The ICeWater Project Approach

Antonio Candelieri

Dante Conti

Davide Cappellini

Francesco Archetti

11th International Conference on Hydroinformatics
HIC 2014, New York City, USA

URBAN WATER DEMAND CHARACTERIZATION AND SHORT-TERM FORECASTING – THE ICEWATER PROJECT APPROACH

CANDELIERI A. (1,2), CONTI D. (2), CAPPELLINI D. (1), ARCHETTI F.(1,2)
(1): Department of Computer Science, Systems and Communication, University of Milano-Bicocca, Italy
(2): Consorzio Milano Ricerche, via Roberto Cozzi 53, Milano, Italy

This paper presents a completely data-driven approach, proposed in the EU-FP7-ICT project ICeWater, which adopts time-series data clustering to identify typical daily urban water demand patterns, and Support Vector Regression for performing reliable short term forecasts. The overall approach has been evaluated on a real case study, that is the urban water demand of the Water Distribution Network (WDN) in Milan, one of the two pilots of ICeWater. The obtained results are promising and the application of the approach also on smart metering data, related to individual customers water consumption, is currently on going. The approach is going to be characterized as a Big Data Analytics solution for supporting Smart Water in modern cities in a more efficient management of the water-energy nexus.

INTRODUCTION

"Understanding *where*, *when* and *why* we use water" [1] is the key to achieve a more sustainable and efficient management of urban Water Distribution Networks (WDNs). The ICeWater project (co-funded by European Commission) aims at integrating robust and proven ICT solutions together with innovative data analytics and decision support functionalities for enabling an efficient management of water-energy nexus.
From a technological point of view, ICT based solutions, such as Supervisory Control And Data Acquisition (SCADA) systems are already widely adopted to monitor and control WDNs, able to report warnings and alarms triggered on specific rules, as well to store data for further analytical approaches, such as advanced functionalities for leakage management [2][3][4].
With respect to water usage and consumption behavior of individual customers, Automatic Metering Readers (AMR) are devices which are gaining new interest in the field of "smart water". Since their deployment involves all the customers of the water utility, AMRs are more expensive than SCADA. However, the availability of huge amount of high-rate consumption data permits to achieve a more accurate customer-segmentation, to define specific demand management strategies and to perform individual demand forecasting.
Although the developed world has been forged on the supply-side, the historical period requires to look to curbing demand as an active, rather than reactive, water management strategy [5][6]. 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.
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, that are:

- the identification of typical water demand patterns, both at aggregated (urban) and individual (customer) level;
- 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 water utility of the ICeWater project.
As result, a reliable short term demand forecasting model has been obtained for the 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 3.1% reduction of energy consumption and 5.2% reduction of energy costs at a WDN in Netherlands.
The proposed *pattern-discovery-based* approach provides a reliable prediction depending on the hourly urban water demand acquired by SCADA at 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 approach (e.g., ARIMA).

METHODS AND MATERIALS

Metropolitana Milanese (MM) is one of the two water utilities of the EU-FP7-ICT project ICeWater. The urban WDN is a highly interconnected infrastructure and is depicted in the following Figure 1. The urban water demand data used in this study has been retrieved from the MM's SCADA, for the period 01 March 2011 to 31 March 2012.

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

Data have been organized into a time-series dataset, where each entry of the dataset is a vector of 24 measurements, that are the hourly volume of water delivered over the day.
As first step, a preliminary preprocessing on the retrieved data has been performed, aimed at identify anomalies and replace missing values. Anyway, this procedure affected only a very limited portion of data due to the reliability of the SCADA system. This is mostly due to the fact that the urban WDN in Milan has a very low leakage level. Respect to this, distortions into the daily urban water demand time-series data are quite rare, making reliable the identification of typical daily consumption patterns.

Time Series Clustering for pattern identification

A specific survey on time-series data clustering is provided in [8]. Commonly used algorithms, performance evaluation, and similarity/dissimilarity measures are also presented. All these considerations are general and relevant, and are taken into account in recent studies related to clustering of time-series data stream [9][10].
A relevant classification of the time-series clustering approaches proposes three different strategies, depending 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 from 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.

A recent and interesting work, which proposes a novel clustering method on time series data [11], summarizes the different types of similarity to compare time-series:

- 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 skill or competence on data analysis. The approach has been designed to address the WDN needs about a more accurate customer profiling, the identification of typical and periodic water consumption behaviors, the continuous monitoring of water usage patterns and variations along time and customers (space).

Forecasting demand through Support Vector 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 has been published in [13]. But also in regression and time series prediction applications, excellent performances were soon obtained [14] contains a more in-depth overview of SVM regression. Additionally, [15] and [16] provide further details on kernels in the context of classification.

DEFINITION OF THE OVERALL 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 in two consecutive phases:

- the former is devoted to cluster 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 forecasting model, based on the Support Vector Regression, able to predict, at one time, the urban water demand for each (remaining) hour of the day, given the consumption at the early morning as acquired by SCADA.

Clustering techniques capturing similarity in shape (i.e., by using cosine 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 to identify 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 consumption data).

As results of this step, a limited set of typical daily urban water demand patterns is identified, where the centroid is select as typical pattern for each cluster.

At the end of this step, a possible relationship between each centroid/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 motivation about the behavior associated to that cluster. To do this, a calendar is depicted, with every day colored according to the corresponding cluster; this supports the evaluation of possible seasonality, surprising periods, daily/weekly habits. It is suitable to take into account at least one year in order to capture possible seasonality.

Successively, each cluster is considered as a dataset and used to train a SVM regression model to predict the urban water demand at a specific hour depending on the first m data. As result, a pool of SVM regression model is generated for each cluster. Thus, the overall number of SVM to be trained is $(24-m) \times K$, where K is the number of identified clusters.

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

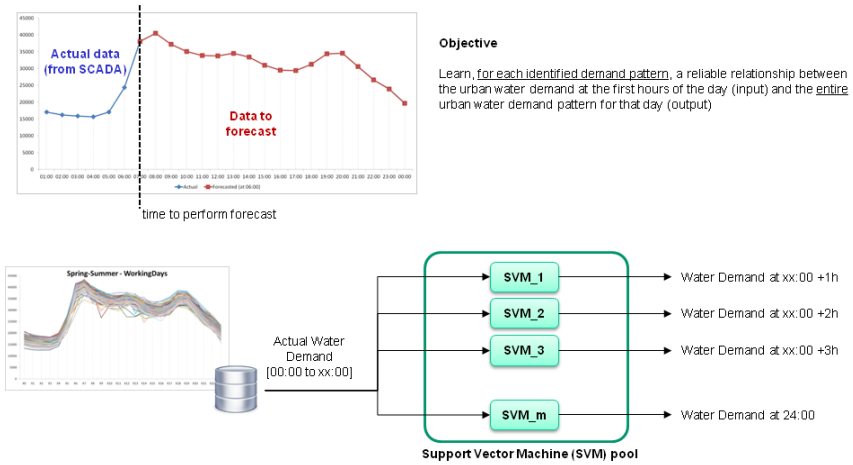

Figure 2. Learning predictive models: one pool of SVM regression models for each typical pattern (centroid of the cluster) identified, one SVM regression model for each hour.

The pools of trained SVM regression models are stored. When demand has to be forecasted, the most suitable pool is identified and retrieved (e.g., according to the expected period of the year and the type of day), then the correspondent models are used to predict the water demand data, at every hour of the day, given the first *m* values acquired through SCADA. The Figure 3 shows this procedure, where demand values at the first 6 hours of the day are considered as input of all the models in the selected pool.

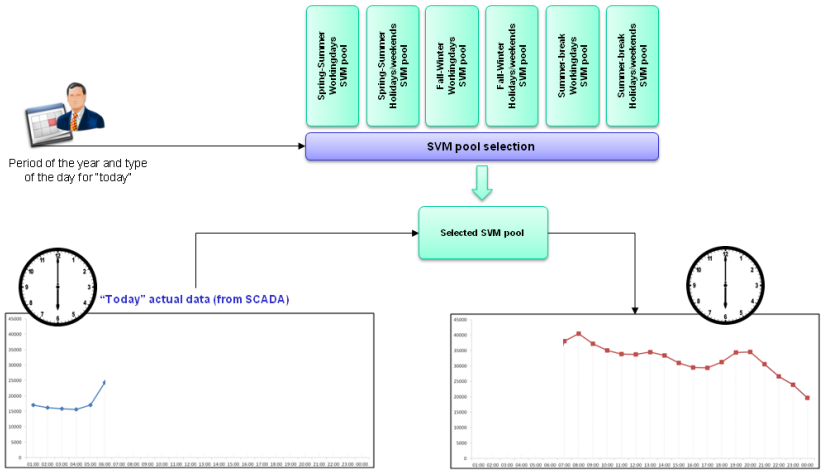

Figure 3. Applying predictive models learned: the most suitable pool of SVM regression models is selected (i.e., depending on period of the year and type of the day) and each model is used to forecast urban water demand at each hour.

RESULTS

In this section the main results are presented. The approach identified 6 typical daily urban water demand patterns on MM's SCADA data, as reported in the following Figure 4.

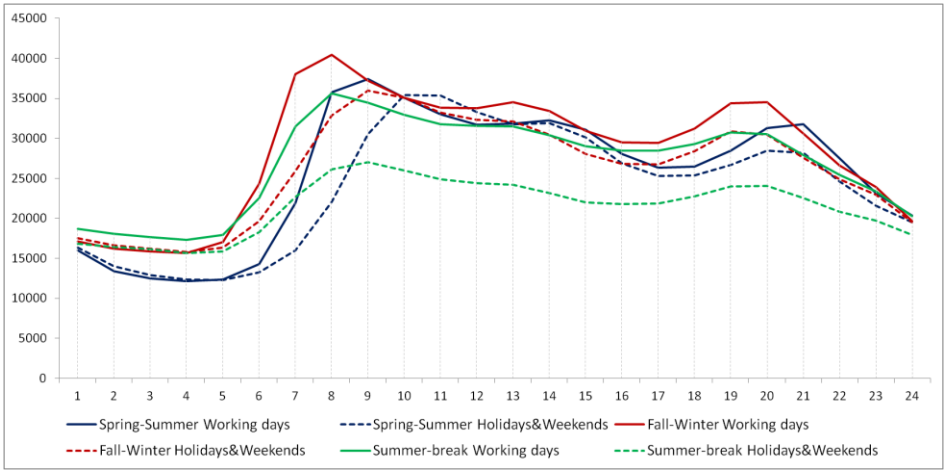

Figure 4. Typical patterns identified in the urban water demand data of the WDN in Milan

By having a look at the following Figure 5 it is possible to identify the association between every daily time-series and the corresponding cluster, over the analyzed time windows. The following relative considerations have been made:

- three different periods of the year have been identified, namely Spring-Summer, Fall-Winter and Summer-break
- 2 different type of day for each time period exist, namely, working-days and holidays-weekends.

Therefore, every cluster is identified by the pair "period of the year" and "type of the day", that is 2 x 3 = 6 clusters.

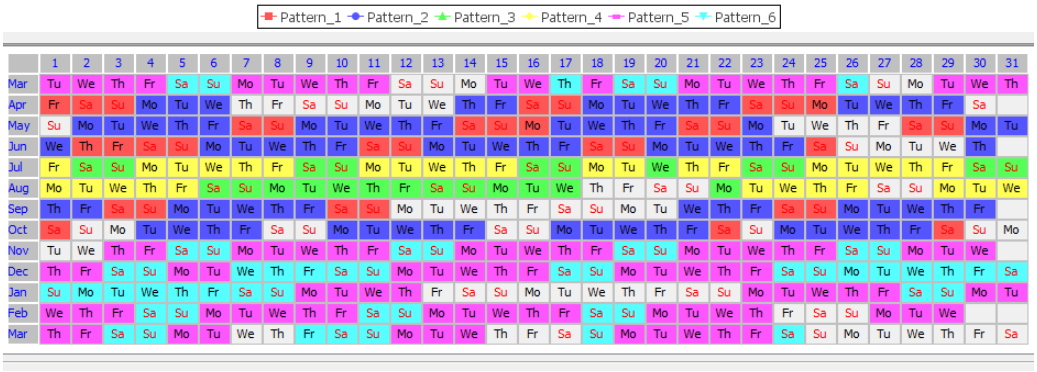

Figure 5. Distribution of the identified clusters/patterns over the analyzed time window

It is really easy to note, having a look at Figure 4, that major differences among the identified typical patterns regard the peaks in consumption in the morning and in the evening.
In particular, the peak in the morning of holidays and week-ends is always delayed of about 1 hour respect to that of working days, for each period of the year.
Moreover, the typical patterns named "Summer-break – working-days" is a really specific daily urban water demand pattern, more "flat" and "low" than the others, and associated to the 15 days in the middle of August, when usually citizens of Milan have their summer holidays and leave the city.

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 urban water 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 final reliability index of the forecasts, the (absolute) percentage error has been computed in correspondence of each hour (i.e., | *actual* – *predicted* | / *actual*). This value has been then averaged on all the "predicted" hours and the result has been used to select the most reliable SVM configurations. Results are reported in the following Table 1, in particular the average and standard deviation of the error for the best and the worst forecasts on 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
Cluster1	0.79%	0.59%	6.11%	2.95%
Cluster2	1.57%	1.18%	14.33%	11.68%
Cluster3	0.84%	0.66%	8.48%	3.53%
Cluster4	1.71%	2.56%	12.84%	7.53%
Cluster5	1.31%	0.93%	7.85%	13.26%
Cluster6	1.10%	0.85%	6.54%	3.46%

CONCLUSIONS

The approach presented in this paper, and developed within the EU-FP7-ICT project ICeWater, proposes a combination of time-series data clustering and Support Vector Machine regression to identify and characterize typical urban water demand patterns and, consequently, provide reliable forecasts in the very short term. This completely data-driven approach has been tested on real data retrieved from the SCADA system of Metropolitana Milanese, the WDN in Milan and one of the two use cases of ICeWater.
The approach has been designed and developed to be applicable both at aggregated level (i.e., urban water demand data from SCADA) and at individual customers level (i.e., consumption data from AMRs); in particular, a study on AMR data is currently on going. To be applicable on customers data, the proposed approach has been developed to be scalable on parallel/distributed architectures, as a Big Data Analytics solution for supporting Smart Water in modern cities.
Finally, it is important to highlight that the identification of typical consumption patterns, in particular when applied at individual users level, enables a more accurate customers-segmentation while supporting the definition of demand management strategies for improving water and costs savings. On the other hand, it enables reliable demand forecasts in the short term which can effectively drive optimization process, such as optimal operations planning to reduce energy-related costs of caption, treatment, storage and distribution.

ACKNOWLEDGEMENTS

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, Vol. 107 No. 50, (2010), pp 21300-21305.*

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

[3] *Candelieri A., Archetti F. and Messina E., "Improving leakage management in urban water distribution networks through data analytics and hydraulic simulation", WIT Transactions on Ecology and the Environment, Vol. 171, (2013), pp 107-117.*

[4] *Candelieri A. and Messina E., "Sectorization and analytical leaks localizations in the H2OLeak project: Clustering-based services for supporting water distribution networks management", Environmental Engineering and Management Journal, Vol. 11 No. 5, (2012), pp 953-962.*

[5] *Hill T. and 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. and Stouffer R.J., "Stationarity is dead: whither water management?", Science, Vol. 319, (2008).*

[7] *Bakker M., Vreeburg J.H.G., Palmen L.J., Sperber V., Bakker G. and 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, Vol. 58 No. 3, (2013), 203-211.*

[8] *Liao T.W., "Clustering of time series data – a survey", Pattern Recognition Vol. 38, (2005), pp 1857-1874.*

[9] *Pereira C.M.M. and de Mello R.F., "TS-stream: clustering time series on data streams", Journal of Intelligent Information Systems, (2013).*

[10] *Kavitha V. and Punithavalli M., "Clustering Time Series Data Stream – A Literature Survey", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 8 No. 1, (2010).*

[11] *Zhang X., Liu J., Du Y., Lv T., "A novel clustering method on time series data", Expert Systems with Applications, Vol. 38, (2011), 11981-11900.*

[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, Vol. 2 No.2, (1998), pp 121–167.*

[14] *Scholkopf B. and Smola A.J., "Learning with Kernels". MIT Press, (2002).*

[15] *Cristianini N. and Shawe-Taylor J., "An Introduction to Support Vector Machines", Cambridge University Press, Cambridge, UK, (2000).*

[16] *Herbrich R., "LearningKernel Classifiers: Theory and Algorithms". MIT Press, (2002).*