Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 89 (2014) 1004 – 1012

Procedia Engineering

www.elsevier.com/locate/procedia

16th Conference on Water Distribution System Analysis, WDSA 2014

Identifying Typical Urban Water Demand Patterns for a Reliable Short-Term Forecasting – The Icewater Project Approach

A. Candelieri[a,b]*, F. Archetti[a,b]

[a]Consorzio Milano Ricerche, Milan 20162, Italy
[b]Department of Computer Science, System and Communication, University of Milano-Bicocca, Milan 20126, Italy

Abstract

This paper presents a computational framework performing, in two stages: urban water demand pattern characterization through time series clustering and reliable hourly water demand forecasting for the entire day based on Support Vector Machine (SVM) regression. An SVM regression model is trained for each cluster identified and for each hour of the day, taking the hourly water demand data acquired at the very first m hours of the day. The approach has been validated on a real case study that is the urban water demand of the Water Distribution Network (WDN) in Milan, managed by Metropolitana Milanese, one of the partner of the EU-FP7-ICT ICeWater project.

Peer-review under responsibility of the Organizing Committee of WDSA 2014

Keywords: urban water demand; forecasting; chategorization; pattern; time series clustering, support vector regression

1. Introduction

Although the developed world has been forged on the supply-side, the historical period requires to look to curbing demand as the main driver for enabling efficient water management strategy [1][2]. 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 distribution by using energy (in particular for pumping) when it is less expensive during the day, while maintaining a satisfactory level of the service [3]. A recent work [4] reports that forecasts led to 3.1% reduction of energy consumption and 5.2% reduction of energy costs at a WDN in Netherlands.

* Corresponding author.
E-mail address: candelieri@milanoricerche.it

Peer-review under responsibility of the Organizing Committee of WDSA 2014
doi:10.1016/j.proeng.2014.11.218

In [5][6] a general overview of recently proposed forecasting models is provided, according to time-scale and the purpose for which the model is predicated.

The analysis of the water demand may reveal important information on the profile, both in spatial and temporal domain, and detect early behavioural changes [7]. Furthermore, the resulting accurate characterization of the water demand is a relevant component in the design of any urban WDN as well the optimization of operations based forecast of urban water demand.

From a technological point of view, many ICT based solutions are already adopted to monitor and control the WDNs, such as Supervisory Control And Data Acquisition (SCADA) systems. These systems gives the opportunity to acquire and store huge quantities of historical data. Although these systems have been mostly used to trigger warnings/alarms according to specific rules and/or to determine optimal pumping schedules [8], they are recently going to be further used to extract values from the collected data [9].

As an example, this data has been widely used, even coupled with simulation software and Machine Learning, to offer innovative leakage management functionalities [10-15]. More recently, this data have been largely used to perform short-term forecast to dynamically optimize pump scheduling under diurnal and seasonal variations of the water demand, in order to maintain a satisfactory level of the service while reducing costs for caption, treatment, storage and distribution.

Another key enabler of smart water is the deployment of Automatic Metering Reader (AMR) at the final consumption point. Since their deployment involves all the customers of the water utility, AMRs are more expensive to deploy than SCADA. Yet they hold the potential of enabling forecasts at individual level.

Irrespectively to the data acquisition system, water demand forecasting may be performed at different time scales according to the specific goal: for operational purposes short-term forecasting for the next 24/48 hours is needed. The output of the model may be one day forecast for general flow control or hourly forecast for detailed pump scheduling and reservoir management. The forecast of the daily demand has been the subject of intensive research; in [16] a comparison among several techniques is presented. Approaches for hourly forecasts, based on similar techniques, have been proposed, generally taking into account one hour as the smallest forecast time step; a relevant and recent exception is reported in [17] where a shorter time interval (e.g., 15 minutes) is indicated as more effective in some cases.

The main contribution of this paper is related to the design and development of a completely data-driven machine learning approach to provide reliable hourly forecasts of the water demand in the short term (24h), with the aim to enable optimal pumping scheduling. The solution is based on two machine learning based analytical stages:

- First stage (unsupervised): clustering time series data related to hourly urban water demand in order to identify typical daily patterns (consumption behaviour)
- Second stage (supervised): learning several of Support Vector Machines (SVM) regression models (i.e., one for each cluster identified at the first stage and for each hour to be forecasted) able to provide reliable predictions.

The input to the water demand forecasting systems proposed in this paper consists of only historical urban water demand data. In [18] it has been already reported that this data may be sufficient to provide fairly accurate forecasts. On the other hand, other solutions exploit the fact that water demand is positively correlated with heat waves and include in their input also weather data [19][20].

The approach has been developed and validated on historical urban water demand data retrieved from the SCADA of Metropolitana Milanese, in Milan, a partner of the EU-FP7-ICT project ICeWater, co-funded by European Commission. The approach proposed in this paper uses the hourly urban water demand data acquired at the first 6 hours of the day and provides, all in once, reliable predictions for the remaining hours of the day. The model does not require explicitly any “on-line updating” and it is not affected by the “time-lag” effect, usually occurring in more classical approach (e.g., ARIMA). As it is completely data-driven, the overall system can be run periodically (e.g., every month) in order to dynamically adapt itself by capturing variations in the consumption behaviour (first stage) and retraining more accurate and up-to-date regression models.

In Section 2, the case study, the data available and the methodological background are described; in section 3 the experimental setting is presented; section 4 presents the relevant results, and a final section provides conclusions.

2. Materials and Methods

2.1. Available data

Metropolitana Milanese (MM) is one of the two water utilities acting as partners and use cases in the EU-FP7-ICT project ICeWater. The urban WDN in Milan (in Fig. 1) has a highly interconnected infrastructure. 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, and is related to more than 5000 customers (buildings) for about 1million of habitants.

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

Data has been organized into a time-series dataset $V = \{v_1, v_2, ..., v_n\}$ consisting of n vectors, that are the days in the observation period, where each v_i is a set of p ordered values, that are the hourly volume of water delivered over the day. As first step, a preliminary pre-processing of the retrieved data has been performed, aimed at identifying anomalies and replacing 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.

2.2. Clustering time series data

Clustering is aimed to partition objects, represented as vectors in a multi-dimensional space, into disjoint groups (clusters), such that some measure of similarity is maximized within groups and minimized between groups. This general aim is valid also in the case objects are time series; the relevant difference is the sequential constraint related to the components of a vector representing a time series (i.e., time-steps). Therefore, data representation and pre-processing, as well definition of a suitable similarity measure, are critical choices to cluster time series data. Two relevant surveys on time series clustering are provided in [21][22].

With respect to data representation and pre-processing, time series clustering strategies may be categorized in:

- working directly with raw data: this approach can be very demanding in the cases vectors in V have high dimensionality or data is affected by autocorrelation or noise;
- working with features (such as average, standard deviation and skewness) extracted by the raw data and then clustering data in the space spanned by these features [23].

- working with a model synthesizing the data, such as ARIMA [24] or Hidden Markov Model [25]; in this case and the clustering is then performed in the space spanned by the parameters of the selected model.

With respect to the similarity between two time series, and at a more qualitative level, the following categorization has been proposed [26]:

- Type 1: similarity in time. The goal is to cluster together series that vary in a similar way on each time step. In this case, data are clustered with respect to the values at time stamps.
- Type 2: similarity in shape. The goal is to cluster together time series having common shape features e.g. common trends occurring at different times or similar sub-patterns.
- Type 3: similarity in change. The goal is to cluster together time series that vary similarly from time step to time step. In this case the data are clustered with respect to the variations between two successive time stamps

More in detail, most of the similarity (or analogously the distance) measures used to cluster time series are those adopted on static data, such as Euclidean distance, Pearson correlation coefficient and cosine distance [27]. Some measures have been specifically proposed for clustering time series, such as Dynamic Time Warping (DTW): given 2 time series DTW aligns the 2 series so that their difference is minimized. Intuitively, the sequences are warped in a nonlinear fashion to match each other, by using dynamic programming in order to identify the best nonlinear warp (i.e., temporal shift).

As daily water consumption patterns are usually strictly dependent on habits and therefore hour of the day, techniques dealing well with temporal shift, such as DTW, are not helpful, while cosine similarity – also known as triangle similarity [26] – handles very well similarity in time, without any time-distortion, thus capturing similar behaviors characterized by peaks and bursts at the same hour of the day.

2.3. Support Vector Regression

Given a dataset *D*, defined as:

$$D=\{(x_i,y_i)\,|\,x_i\in\Re^p, y_i\in\Re\} \text{ with i}=1, ..., \text{n} \tag{1}$$

the basic idea of using SVM [28] for regression [29] consists in searching for a function $f(x)$ that has at most ε deviation from the actually targets y_i for all the data in *D* and, at the same time, is as "flat" as possible. The easiest solution is a linear function in the form:

$$f(x)=\langle w,x\rangle+b \text{ with w}\in\Re^{\text{p}} \text{ and b}\in\Re \tag{2}$$

where <.,.> is the dot product in the *p*-dimensional space. "Flatness" of the solution is represented by small values of *w*. In order to deal with feasibility of the linear solution, another parameter *C* is introduced in the formulation. The regression problem may be thus defined as the following optimization problem:

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + \text{C}\sum_{\text{i=1}}^{\text{n}}\left(\xi_i+\xi_i^*\right)$$

$$\text{subject to} \begin{cases} \text{y}_\text{i}-\langle w,x_i\rangle-b\le\varepsilon+\xi_i \\ \langle w,x_i\rangle+b-y_i\le\varepsilon+\xi_i^* \\ \xi_i,\xi_i^*\ge 0 \end{cases} \tag{3}$$

The constant C > 0 determines the trade-off between the flatness of *f(x)* and the amount up to which deviations larger than ε are tolerated. Only the points outside the shaded region contribute to the cost insofar, as the deviations are penalized in a linear fashion. The best value of *C* is usually chosen through cross validation on training data.

In order to extend Support Vector regression to nonlinear functions, the dual method based on Lagrange multipliers is applied. The Lagrange function *L* is built starting from the original objective function and the corresponding constraints, by introducing a dual set of variables. This function has a saddle point with respect to the primal and dual variables at the solution.

By solving the dual problem the resulting formulation of *f(x)* is usually known as *Support Vector expansion*, because *w* is expressed as a linear combination of the training patterns x_i, making *f(x)* completely independent on the dimensionality *p* of the input: it depends only on the number of Support Vectors (x_i such that α_i are α_i^* are not zero). Moreover, as *f(x)* is described in terms of dot products between data, it is not necessary to compute *w* explicitly, an important consideration to formulate the extension to the nonlinear case.

The simplest method to extend the Support Vector regression to nonlinear data is to preprocess the training set by using a mapping function ϕ, from the original space (Input Space) to some other space (Feature Space) where the linear approach may be successfully applied. The important result is that, rather than explicitly mapping all the data into the new space through the mapping $\phi(x)$, one can use a kernel function. The kernel function enables operations to be performed in the Input Space rather than the Feature Space.

With the "kernel trick" the inner product does not need to be evaluated in the Feature Space. Although this provides a way of addressing the curse of dimensionality, the computation is still critically dependent upon the number of training patterns and to provide a good data distribution for a high dimensional problem will generally require a large training set. Several types of kernel have been proposed (e.g. Polynomial, Radial Basis Functions, Sigmoid, etc.) each one with at least an internal parameter to be set up for mapping [30].

3. Experimental Setting

3.1. Time series clustering for water demand characterization

The approach proposed in ICeWater is characterized by: (i) working directly with the raw data, with all the time series defined in the same time window (i.e., a day) and having the equal length (i.e., 24 data points in the case of consumption data); (ii) working with cosine similarity to capture similarity in time.

More in detail, cosine similarity is given by the cosine of triangle between two vectors, so the range of value of cosine similarity is [-1; 1].

$$s(x,y)=\frac{\langle x,y\rangle}{\|x\|\|y\|} \tag{4}$$

As the components of the urban water demand vectors are not negative, triangle similarity varies in [0; 1]. To perform the time series clustering, a *K*-means algorithm has been adopted, in particular the implementation of the Spherical *K*-means provided by the R package and named "skmeans" [31]. This specific implementation performs a simple *K*-means strategy based on the cosine distance:

$$d(x,y)=1-s(x,y)=1-\frac{\langle x,y\rangle}{\|x\|\|y\|} \tag{5}$$

As a first parameter to be set by the WDN manager/analyst, the overall time period of the observations (daily urban water demand time series) for the analysis has to be provided. Of course, it is suitable to take into account at least one year in order to capture possible seasonality.

The approach works by clustering time-series at two different time levels: at the first one, the daily consumption time-series, averaged for each month, are preliminary clustered; at the second level, a new clustering is performed on

daily consumption time-series belonging to each "first-level" cluster. This two-levels approach usually permits to improve the identification of seasonal behaviors.

The number of clusters is identified via two cluster validity measures (i.e., Calinski-Harabatz and Silhouette, [32]). The distribution of the identified patterns (i.e., resulting time-series computed as centroids of each cluster) has to be visualized over the days of the analyzed period, in order to evaluate possible relations with seasonality, surprising periods, daily/weekly habits.

3.2. Support Vector Regression for Water demand forecasting

At the end of the first stage, a limited set of clusters is identified, where the centroids are select as typical pattern for each cluster. To perform the second stage, each cluster is considered as a dataset. The first m columns are the input variables, and they correspond to the urban water demand values observed at the first m hours of the day. The output variable, to be predicted, is the j-th column of the original dataset, with $j=m+1, \ldots, 24$. For each j, a SVM regression model is trained, taking as input only the first m columns of the original dataset.

The pools of trained SVM regression models are stored. When a new pattern of m daily demand values arrives a pool of SVMs is selected. In this study the retrieval is performed according to the period of the year and the type of day, since each identified pattern has been associated to daily time-series belonging to the corresponding cluster.

The selected pool of SVMs are thus used to predict the water demand data at the remaining 24-m hours.

With respect to the issue of model calibration, the entire approach has been developed in order to automatically capture changes in the behaviors and adapt the forecasting model. The re-learning of the entire system is performed at very low frequency, such as every month, and by taking into account at the least data of the latest year. To perform the updating of the system, both the stages have to be executed: the identification of typical consumption patterns (time series clustering) and the training of the pools of SVM regression models.

The proposed solution is completely data-driven and is able to automatically capture modifications of the typical water consumption behaviors, including changes to the number/shape of typical urban demand pattern.

4. Results

4.1. Urban water demand characterization

The bi-level clustering approach identified 6 typical daily urban water demand patterns on MM's SCADA data, as reported in the following Fig. 2.

By looking at the following Fig. 3 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 (named "Spring-Summer", "Fall-Winter" and "Summer-break");
- two different types of day for each time period exist (named "working-days" and "holidays-weekends").

Every cluster is identified by the pair "period of the year" and "type of the day" (i.e. 2 x 3 = 6 clusters).

It is really easy to note, by looking at Fig. 7, 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.

4.2. Urban water demand forecasting

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 (m=6). 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.

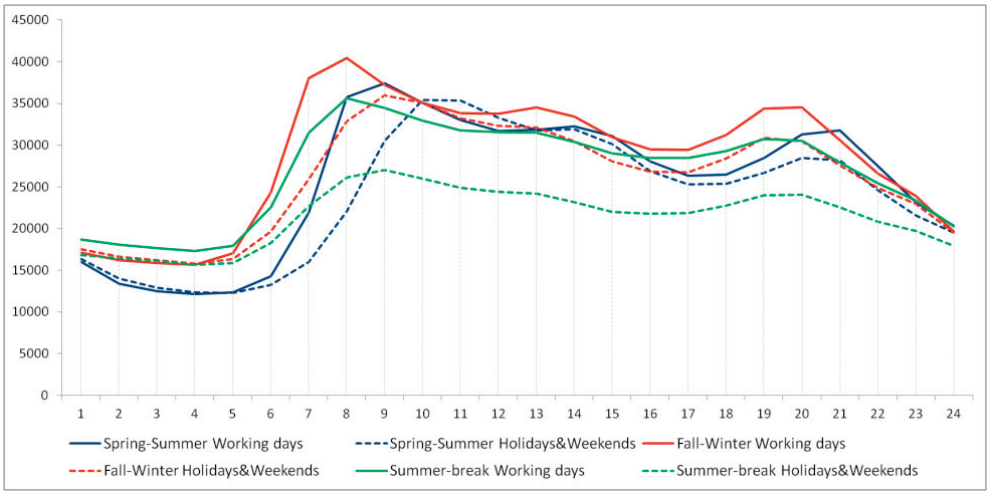

Fig. 2. Typical patterns identified in the urban water demand data of the WDN in Milan.

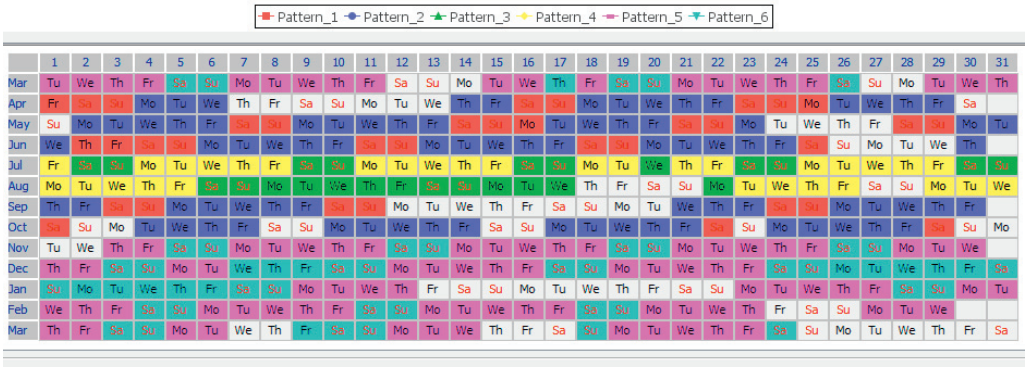

Fig. 3. Distribution of the identified clusters/patterns over the analyzed time window.

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 unsigned relative error has been computed in correspondence of each hour:

$$E_i^j = \frac{\left| p_i^j - a_i^j \right|}{a_i^j} \tag{6}$$

where $p_i{}^j$ and $a_i{}^j$ are, respectively, the predicted and the actual value of demand for hour i and time-series j. As overall error on the time series j, the average is computed:

$$\overline{E}^j = \frac{100}{24-m} \cdot \sum_{i=m+1}^{24} E_i^j \tag{7}$$

which is usually known as Mean Absolute Percentage Error (MAPE) [6].

Finally, as results the best (minimum MAPE) and worst (maximum MAPE) forecasts, for each cluster, have been reported in the following Table 1:

Table 1. Error (MAPE) for the best and the worst forecasts in each cluster with standard deviation.

	Best		Worst	
	Mean	Std Dev	Mean	Std Dev
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%

5. Conclusions

The approach presented in this paper has been developed and validated in the EU-FP7-ICT project ICeWater. Unsupervised (time series clustering) and supervised (Support Vector regression) machine learning strategies are combined in this two-stages framework in order to identify typical urban water demand patterns and successively provide reliable short term forecasts for each hour of the day.

The proposed approach has been tested on real data stored into the SCADA system of Metropolitana Milanese, the WDN in Milan and one of the two use cases of ICeWater, and has been designed 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 in Milan.

Finally, it is important to remark that the identification of typical consumption patterns, enables reliable demand forecasts in which can effectively drive process optimization, such as optimal operations planning to reduce energy-related costs of caption, treatment, storage and distribution and, support the definition of demand management strategies and in particular, when applied at individual users data, an accurate customers-segmentation and service.

The entire system has been developed to be dynamically updated. The re-training is performed periodically (e.g., every month) and by taking into account at the least data of the latest year. Both the stages have to be executed: the identification of typical consumption patterns (time series clustering), with the aim to capture modifications of the typical water consumption behaviors, including changes to the number/shape of typical urban demand pattern, and the training of the pools of SVM regression models, with the aim to provide reliable forecasts.

Acknowledgements

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

References

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

[2] P.C.D. Milly, J. Betancourt, M. Falkenmark, R.M. Hirsch, Z.W. Kundzewicz, D.P. Lettenmaier, R.J. Stouffer, Stationarity is dead: whither water management?, Science, 319 (2008).

[3] T.G. Mamo T.G., J. Ilan, S. Isam, Urban Water Demand Forecasting Using the Stochastic Nature of Short Term Historical Water Demand and supply Pattern, Journal of Water Resource and Hydraulic Engineering, 2 (2013) 92-103.

[4] M. Bakker, J.H.G. Vreeburg, L.J. Palmen, V. Sperber, G. Bakker, L.C. Rietveld, 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), (2013), 203-211.

[5] L.A. House-Peters, H. Chang, Urban water demand modeling: review of concepts, methods, and organizing principles, Water Resources Research 47 (2011) W05401.

[6] E. Donkor, T. Mazzuchi, R. Soyer, J. Roberson, Urban water demand forecasting: a review of methods and models, Journal of Water Resources Planning and Management, 140 (2012) 146–159.
[7] J. Gama, Knowledge discovery from data streams, 1st Edn. Chapman & Hall / CRC (2010).
[8] S.M. Bunn, L. Reynolds, The energy-efficiency benefits of pumps-scheduling optimization for potable water supplies, IBM Journal of Research and Development 53 (2009) 1-13.
[9] K. Hassaballah, A. Jonoski, I. Popescu, D.P. Solomatine, Model-based optimisation of downstream impact during filling of a new reservoir: case study of Mandaya/Roseires reservoirs on the Blue Nile River, Water Resources Management, 26 (2012) 273-293.
[10] L. Xia, L. Guo-jin, Leak detection of municipal water supply network based on the cluster-analysis and fuzzy pattern recognition, International Conference on E-Product E-Service and E-Entertainment (ICEEE), 1 (2010) 7-9.
[11] A. Nasir, B.H. Soong, S. Ramachandran, Framework of WSN based human centric cyber physical in-pipe water monitoring system, Proc. 11th International Conference on Control, Automation, Robotics and Vision, (2010) 1257-1261.
[12] W. Lijuan, Z. Hongwei, J. Hui, A Leak Detection Method Based on EPANET and Genetic Algorithm in Water Distribution Systems, Software Engineering and Knowledge Engineering: Theory and Practice – Advances in Intelligent and Soft Computing, 14 (2012) 459-465.
[13] A. Candelieri, 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, 11 (2012) 953-962.
[14] A. Candelieri, F. Archetti, E. Messina, Improving leakage management in urban water distribution networks through data analytics and hydraulic simulation, WIT Transactions on Ecology and the Environment, 171 (2013) 107-117.
[15] A. Candelieri, D. Conti, F. Archetti, 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).
[16] J. Adamowski, H. Fung Chan, S.O. Prasher, B. Ozga-Zielinski, A. Sliusarieva, Comparison of multiple linear and nonlinear regression, autoregressive integratedmoving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. Water Resources Research, 48 (2012) W01528.
[17] M. Bakker, J.H.G Vreeburg, L.J. Palmen, V. Sperber, G. Bakker, L.C. Rietveld, Better water quality and higher energy efficiency by using model predictive flow control at water supply systems. Journal of Water Supply: Research and Technology e AQUA 62 (2013) 1-13.
[18] S. Alvisi, M. Franchini, A. Marinelli, A short-term, pattern-based model for water-demand forecasting. Journal of Hydroinformatics 9 (2007) 39-50.
[19] G. Ghiassi, D.K. Zimbra, H. Saidane, Urban water demand forecasting with a dynamic artificial neural network model. Journal of Water Resources Planning and Management 134 (2008) 138-146.
[20] M. Herrera, L. Torgo, J. Izquierdo, R. Pérez-García, Predictive models for forecasting hourly urban water demand. Journal of Hydrology 387 (2010) 141-150.
[21] T.W Liao, Clustering of time series data – a survey, Pattern Recognition 38 (2005) 1857-1874.
[22] V. Kavitha, M. Punithavalli, Clustering Time Series Data Stream – A Literature Survey, (IJCSIS) International Journal of Computer Science and Information Security 8 (2010) 289-294.
[23] T. Räsänen, M. Kolehmainen, Feature-based clustering for electricity use time series data, Adaptive and Natural Computing Algorithms, Springer (2009).
[24] M. Corduas, D. Piccoloa, Time series clustering and classification by the autoregressive metric, Computational Statistics and Data Analysis, 52 (2008) 1860–1872.
[25] M. Bicego, V. Murino, M.A.T. Figueiredo, Similarity-based clustering of sequences using hidden Markov models. In Lecture Notes in Computer Science 2734 (2003) 95–104.
[26] X. Zhang, J. Liu, Y. Du, T. Lv, A novel clustering method on time series data, Expert Systems with Applications, 38 (2011) 11981-11900.
[27] C.M.M. Pereira, R.F. de Mello, TS-stream: clustering time series on data streams, Journal of Intelligent Information Systems, (2013).
[28] V. Vapnik, The Nature of Statistical Learning Theory. Springer, New York, 1995.
[29] V. Vapnik, Statistical Learning Theory. New York, Wiley, 1998.
[30] B. Scholkopf, A.J. Smola, Learning with kernels. Support Vector Machines, regularization, optimization and beyond, MIT Press, USA (2002).
[31] R. Maitra, I.P. Ramler, A k-mean-directions algorithm for fast clustering of data on the sphere. Journal of Computational and Graphical Statistics, 19 (2010) 377–396.
[32] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. Pérez, I. Perona, An extensive comparative study of cluster validity indices, Pattern Recognition, 46 (2013) 243-256.