water

MDPI

Article

Clustering and Support Vector Regression for Water Demand Forecasting and Anomaly Detection

Antonio Candelieri

Department of Computer Science, Systems and Communication, University of Milano-Bicocca, viale Sarca 336, 20126 Milan, Italy; antonio.candelieri@unimib.it or candelieriantonio@gmail.com; Tel.: +39-339-4652297

Academic Editor: Ataur Rahman
Received: 2 February 2017; Accepted: 10 March 2017; Published: 18 March 2017

Abstract: This paper presents a completely data-driven and machine-learning-based approach, in two stages, to first characterize and then forecast hourly water demand in the short term with applications of two different data sources: urban water demand (SCADA data) and individual customer water consumption (AMR data). In the first case, reliable forecasting can be used to optimize operations, particularly the pumping schedule, in order to reduce energy-related costs, while in the second case, the comparison between forecast and actual values may support the online detection of anomalies, such as smart meter faults, fraud or possible cyber-physical attacks. Results are presented for a real case: the water distribution network in Milan.

Keywords: time-series clustering; support vector regression; water demand forecasting; anomaly detection

1. Introduction

Short-term water demand forecasting is a key input for pumping schedule optimization to guarantee a satisfactory level of service while reducing costs for capture, treatment, storage and distribution by using energy when it is cheaper during the day [1] and encouraging water savings. According to a recent study [2], water demand forecasts enabled a 3.1% reduction of energy consumption and a 5.2% reduction of energy costs at a Water Distribution Network (WDN) in the Netherlands.

More recently, the deployment of Automatic Meter Reading (AMR) and Advanced Metering Infrastructures (AMI) has enabled innovative added-value services, such as real-time monitoring and customer profiling, real-time leakage detection, predictive simulation, anomaly detection, and fraud detection (such as False Data Injection, FDI).

With respect to innovative services for the individual customer, an example is provided by New York City [3], where 834,000 customers since January 2012 have had wireless meters and real-time online access to water bills. The total cost of the citywide AMR system installation was approximately $250 million, but it is reducing the cost of City government by substantially reducing billing disputes and other costly aspects of the quarterly billing system.

This paper presents a new short-term forecasting approach (24-h horizon) with hourly periodicity (even a shorter time scale can be handled) designed and developed to be:

- completely data-driven; it considers as input only historical water demand data;
- completely independent of the data source and therefore directly applicable to urban water demand (SCADA data) as well as individual customer consumption (AMR data);
- based on two-stage learning: (i) identifying and characterizing typical daily consumption patterns and (ii) dynamically generating a set of forecasting models for each typical pattern identified in the previous stage. This approach deals with the nonlinear variability of water demand at different levels, automatically characterizing periodicity (e.g., seasonality) and behaviour-related differences of different types of days and hours of the day;

- able to provide reliable forecasts of the urban water demand (SCADA data) in the short term in order to support optimization of operations, in particular pump schedule optimization;
- able to detect possible anomalies in typical water consumption behaviour at the individual customer level (AMR data) associated with metering faults, possible frauds and cyber-physical threats.

This approach has been validated on urban water demand data acquired through the SCADA system of the Metropolitana Milanese, the urban water distribution utility in Milan, Italy, and a set of 26 AMRs (individual consumption) installed in a small area within a Pressure Management Zone (PMZ) named "Abbiategrasso" in the Matherial section. It is important to highlight that the AMRs are installed "per building", a common situation in Italy, so individual customers are not individual households.

State of the Art of Water Demand Forecasting

A plethora of water demand forecasting approaches has been proposed with different characteristics associated with specific goals, forecast horizons, periodicity of the data, and variables used. The availability and selection of specific variables can guide—as well as limit—the identification of the approach to adopt. Usually, those dealing with variables that can be easily collected, monitored, and used by the water utility (i.e., SCADA and AMR data) are preferred, reducing the risk of adding noise/errors from "external" data/information sources as well as avoiding connection of the automation system to the internet for ICT security reasons (e.g., weather forecast services) [4].

A recent study [5] provides a meta-analysis of the empirical literature on water demand forecasting to identify explanations of cross-study variations in the accuracy of different approaches and reporting. It concludes that forecasting depends on relevant characteristics such as demand periodicity, modelling method, forecasting horizon, model specification, and sample size.

A more technical and relevant overview of urban water demand forecasting is provided in [6], where a categorization of the approaches is proposed with respect to a set of further characteristics, including the difference between linear and nonlinear methods [7] (since linear methods are usually not as effective due to the intrinsic nonlinearity of water demand data), as well as between modelling and predicting time series data [8] (where the former is devoted to identify periodicity, such as seasonality, as trends, while the latter usually uses short memory data along with a model of the underlying data generation process to provide predictions).

A relevant research track is about modelling, simulating and predicting demand, in particular with respect to individual customer and end-use. Residential as well as non-residential water demand is one of the most difficult parameters to be estimated when modelling drinking water distribution networks. Several modelling approaches have been proposed, mainly stochastic [9,10], also for small time-scale and small spatial cases, such as Neyman-Scott Rectangular Pulse (NSRP) [11], SIMDEUM—an end-use model to predict water demand at small time and spatial scale according to statistical information about users and end-use [12], a time-dependent Markovian queueing system [13], Overall Pulse (OP)—a system allowing for the generation of the overall domestic demand as displayed at the house water meter [14].

Recent relevant advances have been achieved through the adoption of machine learning for the implementation of effective short-term water demand forecasting, as well as in hydraulic engineering issues, in general [15,16]. Although Artificial Neural Networks (ANNs) are the first and most widely adopted techniques [17–21], the number of works using other promising strategies, such as Support Vector Machine (SVM) regression [22–24], has been increasing. In particular, as reported in [20], SVM regression proved to be the best choice to implement hourly water demand forecasting when compared with ANNs, Projection Pursuit Regressions (PPR), Multivariate Adaptive Regression Splines (MARS), Random Forests and weighted pattern-based water demand forecasting. SVM regression proved to be effective for real-time [25] and dynamic forecasting [26], and online Multiple Kernel Learning updates the water demand forecasting model while improving accuracy through the combination of different kernel functions [22].

To identify the optimal tuning of a water demand forecasting model, several approaches based on meta-heuristics (e.g., Genetic Programming, Genetic Algorithms, etc.) have been proposed in order to efficiently explore the space of possible configurations; some examples are Evolutionary Artificial Neural Networks (EANN) [15], Teaching-learning-based optimization (TLBO) [27] and a combination of phase space reconstruction—used to feed the determinants of water demand with proper lag times—called GEP (Genetic Expression Programming) and SVM [28].

With respect to the idea of performing the two consecutive stages of analysis proposed in this paper, a similar solution was suggested to forecast energy consumption: in that case, the forecasting procedure was driven by a preliminary clustering of pattern sequences (i.e., energy consumption time-series) [29], and more recently, a general algorithm for pattern sequence-based forecasting has been released [30]. The main difference is in generating the forecasts: in [29,30], the overall data stream is transformed into a sequence of cluster assignments with a cluster label for each day; the forecast is computed by averaging the daily consumption patterns, which occur after a given sub-sequence of cluster labels, while the approach proposed in this paper uses the results of clustering to perform a supervised learning stage inferring a number of forecasting models—one for each cluster and for each hourly consumption to be predicted.

Preliminary results of the proposed approach have been presented in [31,32], where the focus was solely on forecasting, even for AMR data, and no anomaly detection was addressed. With respect to the urban water demand data, this paper presents results on a more recent set of data than that used in the preliminary study [31].

2. Materials and Methods

This section presents the two different sets of data used in this study to validate the proposed water demand forecasting approach. Both the datasets are related to the same urban water network in Milan, Italy.

2.1. SCADA Data

The first set of data is related to urban water demand data collected through the SCADA system during the period 1 October 2012 to 30 September 2013 for more than 5000 customers (buildings) serving approximately 1 million habitants.

Most relevant characteristics of the water distribution network are as follows:

- 149,639 junctions
- 118,950 pipes
- 26 pumping stations
- 501 wells and well pumps
- 33 storage tanks
- 95 booster pumps
- 36,295 valves
- 602 check valves
- total base demand 7.5 ± 4.2 m^3/s.

The water distribution network is shown in Figure 1.

To perform the analysis, the hourly water demand data was organized into a time-series dataset $D = \{x_1, x_2, \ldots, x_n\}$ consisting of n vectors, one for each day in the observation period, where each vector x_i is a set of 24 ordered values that are the hourly volume of water delivered in the i-th day. As a first step, a preliminary pre-processing of the retrieved data was performed to evaluate data quality with respect to outliers and missing values.

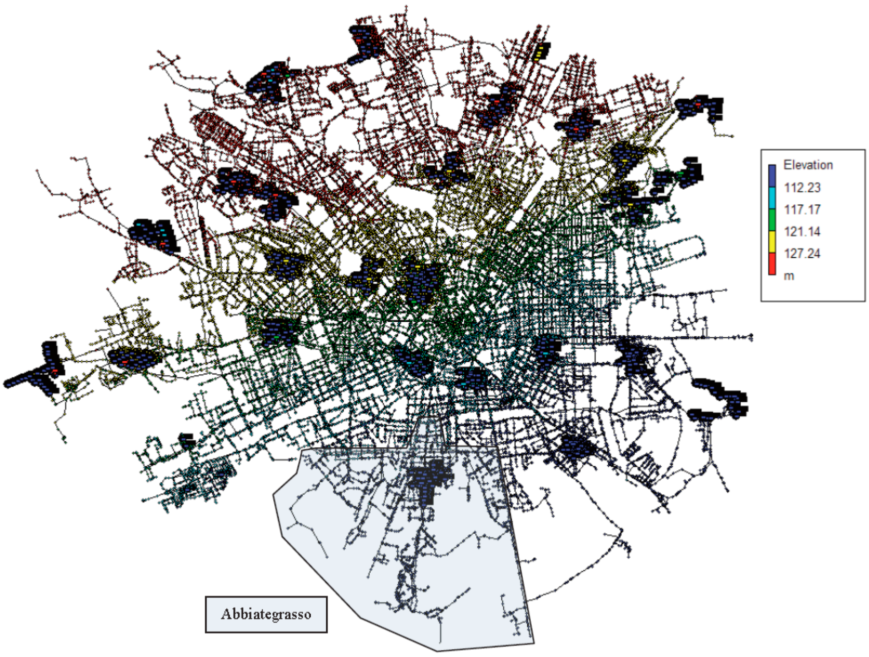


Figure 1. The urban water distribution network in Milan. Highlighted, in the South, the Pressure Management Zone (PMZ) named "Abbiategrasso", where a pilot zone was identified for the installation of Automatic Meter Reading (AMRs).

2.2. AMR Data

The set of available data is related to the individual water consumption values collected in the period September–December 2014 following the installation of AMRs in a limited pilot area within the Pressure Management Zone (PMZ) named "Abbiategrasso". For every AMR, a time-series dataset has been defined: $D^j = \{x^j_1, x^j_2, \ldots, x^j_n\}$, where j is the index identifying the j-th AMR and n is the number of days available for that AMR. Even for AMR data, each x^j_i, $i = 1, \ldots, 24$ in the dataset is a 24-dimensional vector where every component is the hourly water consumption value. The available set on AMR data is small in that it was a piloting activity performed in the EU project ICeWater, and it consisted of 26 AMRs with 110 vectors of 24 hourly water consumption values in each one. Each AMR is devoted to collecting the consumption data of a customer, which in the case of Milan is a building, so different types of patterns can be observed according to residential, non-residential and mixed types of buildings. In particular, 19, 5, and 2 out of the 26 AMRs are associated with residential, non-residential and mixed water usage patterns, respectively.

2.3. Time Series Clustering

The general goal of any clustering algorithm is to group objects, represented as vectors in a multi-dimensional space, such that some measure of similarity is maximized within groups and minimized between groups. Although this general goal is still valid for time-series data clustering the sequential nature of this type of data requires specific choices with respect to data representation, pre-processing, and similarity measure. An extended overview on time-series data clustering is provided in two recent papers [33,34].

With respect to data representation, and according to the idea proposed in this paper, the choice was to work directly with the raw data of the 24-dimensional vectors defined in the previous section. This choice can be very demanding in cases where vectors have high dimensionality, but that is not the case in this study. It allows for fine-grain properties that could be otherwise masked by simply approximating each time series through a set of features (e.g., mean, median, standard deviation, kurtosis and skewness) as well as through a model (e.g., ARIMA).

For the definition of suitable similarity measures, the categorization proposed in [35] was considered:

- Type 1: similarity in time. The goal is to cluster together series that vary in a similar way at each time step. In this case, time series can be clustered by capturing repetitive behaviours occurring always at the same time step or in the same time window (e.g., peak/burst hours).
- 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.

In particular, the paper [35] combines two different similarity measures in comparing two time series: the triangle similarity (aka cosine similarity) to measure similarity in time and Dynamic Time Warping (DTW), a measure specifically defined for time-series data in order to compute similarity in shape.

The approach proposed in the paper is based on the idea that time-series clustering can capture typical consumption behaviours characterized by recurrent peak/burst hours depending on water consumption habits. Similarity in time measures is more suitable in capturing classes of typical behaviours, and cosine similarity was chosen for implementation of the proposed approach.

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

$$s(x_i, x_j) = \frac{\langle x_i, x_j \rangle}{\|x_i\| \, \|x_j\|}$$

As the components of the urban water demand vectors are not negative, triangle similarity may vary from [0 to 1]. The spherical *k*-means algorithm provided by the R package "skmeans" [36] is used, which implements a simple *k*-means strategy based on the cosine distance:

$$(x_i, x_j) = 1 - s(x_i, x_j) = 1 - \frac{\langle x_i, x_j \rangle}{\|x_i\| \, \|x_j\|}$$

According to the clustering algorithm used, the number of clusters must be identified in order to identify a suitable set of different patterns representing typical water consumption behaviours. To select the most suitable number of clusters *k*, two cluster validity measures have been used that are referenced by Silhouette and Calinski-Harabatz [37].

Silhouette is defined as follows:

$$Sil(C) = \frac{1}{n} \sum_{C_k \in C} \sum_{x_i \in C_k} \frac{b(x_i, C_k) - a(x_i, C_k)}{\max\{a(x_i, C_k), b(x_i, C_k)\}}$$

where

$$a(x_i, C_k) = \frac{1}{|C_k|} \sum_{x_j \in C_k} d(x_i, x_j)$$

$$b(x_i, C_k) = \min_{C_l \in C \setminus C_k} \left\{ \frac{1}{|C_l|} \sum_{x_j \in C_l} d(x_i, x_j) \right\}$$

x_i is the *i*-th object of the dataset, that is a 24-dimensional time series of the hourly water consumption for a specific day.

C_k is the *k*-th cluster.

C is the set of clusters identified.

$d(x_i, x_j)$ is the distance between two objects that are two time series.
n is the overall number of objects in the dataset, that is, the number of daily time-series available.
Calinski-Harabatz is defined as follows:

$$CH(C) = \frac{n - |C|}{|C| - 1} \frac{\sum_{C_k \in C} |C_k| d(\overline{C}_k, \overline{X})}{\sum_{C_k \in C} \sum_{x_i \in C_k} d(x_i, \overline{C}_k)}$$

where

$\overline{C}_k$ is the centroid of the k-th cluster.
$\overline{X}$ is the mean vector of the whole dataset.

At the end of the clustering procedure, the centroid of each cluster is selected as the representative water demand pattern for all the time-series data belonging to that cluster, and, consequently, every day in the analysed period can be associated with only one typical consumption behaviour (i.e., centroid). Visualizing this "day-centroid" association over the observation period, it is possible to identify seasonality, surprising periods, and daily/weekly habits. To make this kind of consideration consistent, the overall time period to be analysed should be at least one year, in particular with respect to the identification of seasonality. To make possible seasonality more evident, the clustering procedure is applied to a two-level schema. A new dataset is created from the original one by computing the average consumption pattern for each month:

$$z_m = \frac{1}{N_m} \sum_{i=1,\ldots,N_m} x_i$$

where m is the m-th and N_m is the number of days in that month. Thus, the new dataset consists of M time-series data where M is the number of months. A first clustering is performed on this new dataset in order to identify k_1 clusters corresponding to seasonality (i.e., months characterized by similar average daily patterns). Cluster assignment at this first level is used to label the original time-series dataset; then, according to the attached labels, k_1 sub-datasets are selected from the original dataset in order to perform clustering on each of them. The best ${k_2}^q$ is selected for each sub-dataset, with $q = 1, \ldots, k_1$. A schematic representation of the process in provided in Figure 2.

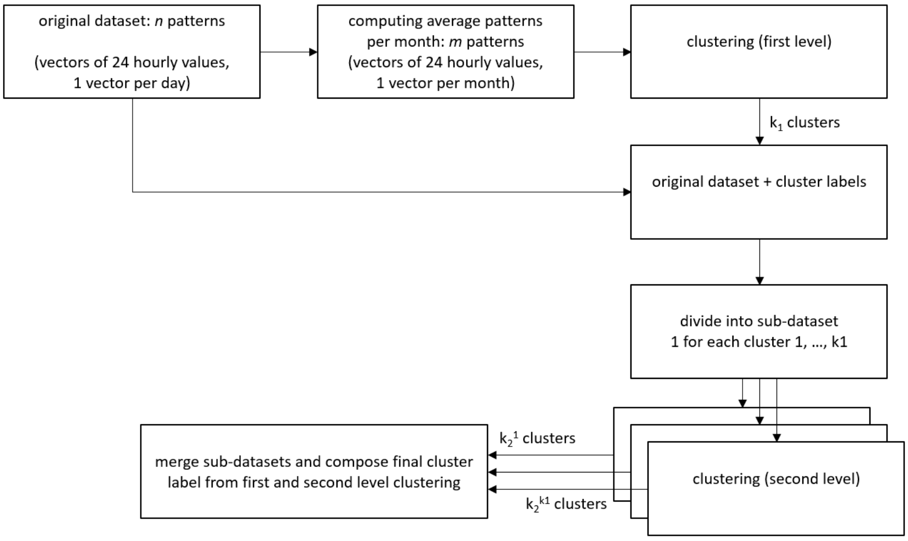

original dataset: *n* patterns
(vectors of 24 hourly values, 1 vector per day)
computing average patterns per month: *m* patterns (vectors of 24 hourly values, 1 vector per month)
clustering (first level)
k1 clusters
original dataset + cluster labels
divide into sub-dataset 1 for each cluster 1, …, k1
clustering (second level)
k2^1 clusters
k2^k1 clusters
merge sub-datasets and compose final cluster label from first and second level clustering

Figure 2. A schematic representation on the two-level clustering procedure that is the first stage of the proposed approach.

2.4. Support Vector Regression Based Demand Forecasting

Given a dataset D, defined as:

$$D = \left\{ \left(x^i, y^i \right) \middle| x^i \in \mathbb{R}^p, y^i \in \mathbb{R} \right\}$$

With $i = 1, \ldots, n$, the basic idea of using SVM [38] for regression [39] consists of searching for a function $f(x)$ that has at most ε deviations from the actual 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:

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

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

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

$$\text{subject to } \begin{cases} \mathrm{y}^{\mathrm{i}} - \langle w, x^i \rangle - b \leq \varepsilon + \xi^i \\ \langle w, x^i \rangle + b - y^i \leq \varepsilon + \xi^{i*} \\ \xi^i, \xi^{i*} \geq 0 \end{cases}$$

The constant $C > 0$ determines the trade-off between the flatness of $f(x)$ and the amount 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 the cross-validation of training data.

To solve the optimization problem above, the dualization 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 the Support Vector expansion because w is expressed as a linear combination of the training patterns x^i, making $f(x)$ completely independent of the dimensionality p of the input; it depends only on the number of Support Vectors (x^i such that the associated Lagrangian multiplier is not zero). As $f(x)$ is described in terms of dot products between data, it is not necessary to compute w explicitly, an important consideration when formulating the extension to the nonlinear case.

The simplest method to extend the Support Vector regression to nonlinear data is to pre-process 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 $\varphi(x)$, one can use a kernel function. The kernel function enables operations to be performed in the Input Space rather than the Feature Space. 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 tuned [40].

The idea proposed in this paper is to learn a number of SVM regression models to perform forecasting. At the end of the clustering procedure, a limited set of clusters is identified; each one of them is considered a dataset in this second stage of the proposed approach. The first p components of each vector of a cluster are the input variables of the SVM regression model and correspond to the hourly water consumption of the first p hours of the day. The target variable to predict is the h-th column of the original dataset, with $h = p + 1, \ldots, 24$; a SVM regression model is trained for each h, and all of them form a "pool" of SVMs for the specific cluster.

The following Figure 3 summarizes this second stage of the proposed approach.

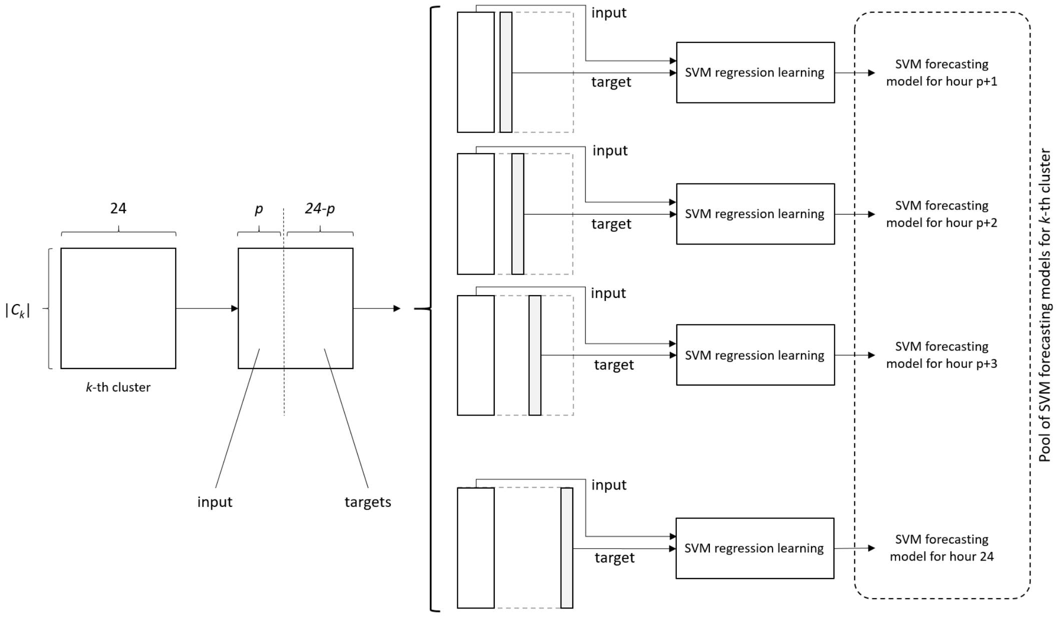

24
p
24-p
$|C_k|$
k-th cluster
input
targets
input
target
SVM regression learning
SVM forecasting model for hour p+1
input
target
SVM regression learning
SVM forecasting model for hour p+2
input
target
SVM regression learning
SVM forecasting model for hour p+3
input
target
SVM regression learning
SVM forecasting model for hour 24
Pool of SVM forecasting models for k-th cluster

Figure 3. A schematic representation of the second stage of the approach: learning a pool of Support Vector Machine (SVM)-based regression models for each cluster to predict hourly water demand.

Measuring forecast errors is crucial for the selection of the models' parameters as well as to monitor the accuracy and reliability of the generated forecasts; degradation of performance may require an updating of the models. The basic idea of any error measure consists of comparing forecasts with observations; several measures have been proposed, but the most widely adopted in the field of water demand forecasting is the Mean Absolute Percentage Error (MAPE) [2,7,22–24,26]. This measure is denoted with:

Y—time-series of observed water demand (at any forecast periodicity),
Y_t—water demand observed at the time t,
$\hat{Y}$—time-series of forecasted water demand (at any forecast periodicity),
$\hat{Y}_t$—water demand forecasted at the time t, and
N—time-series length;

Then, MAPE is computed as the average of the absolute values of the difference, in percentage, between the forecasted and observed data at each time step:

$$\text{MAPE} = \frac{100}{N} \sum_{t=1}^{N} \left| \frac{Y_t - \hat{Y}_t}{Y_t} \right|$$

It is important to highlight [3] that MAPE might be the only error measure that can be used to compare forecasting performance among different utilities because it is independent of system capacity and independent of the unit measure.

Finally, to use the learned models to provide forecasts, the specific SVM pool to use is retrieved according to the time of year and the type of day, that is, information resulting from the clustering procedure performed in the first stage. When a new sequence of p hourly demand values is available, it is used as input to the selected SVM pool, and each SVM regression model in the pool provides the predicted value for the associated hour of the 24-p remaining hours of the day. A schematic representation of the process is provided in Figure 4.

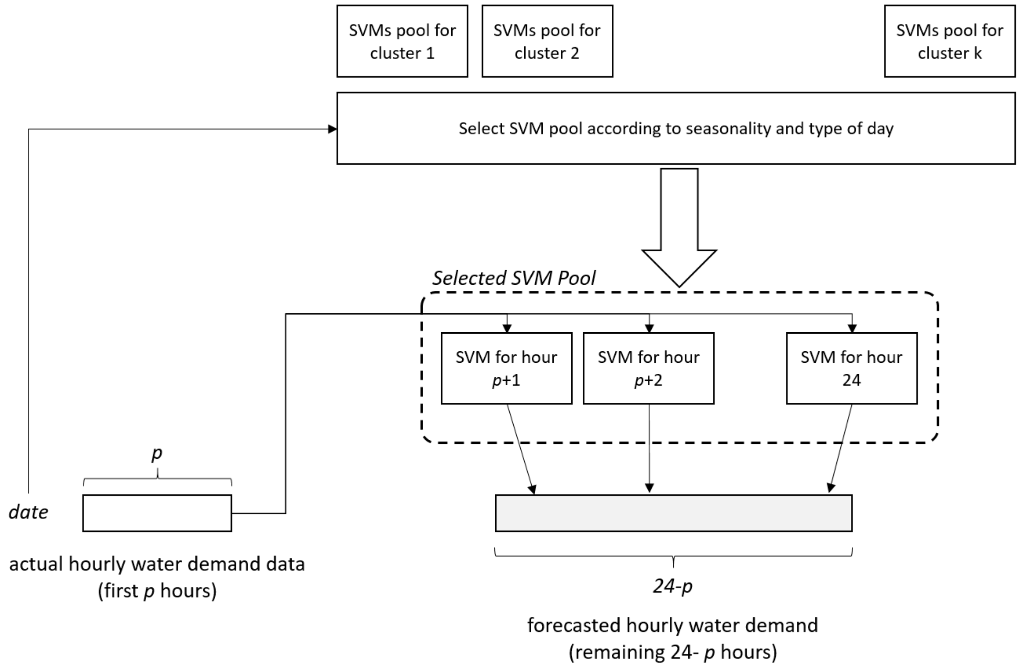

SVMs pool for cluster 1
SVMs pool for cluster 2
SVMs pool for cluster k
Select SVM pool according to seasonality and type of day
Selected SVM Pool
SVM for hour p+1
SVM for hour p+2
SVM for hour 24
p
date
actual hourly water demand data (first p hours)
24-p
forecasted hourly water demand (remaining 24- p hours)

Figure 4. A schematic representation of how forecasting is generated online according to the results obtained from the two stages of the approach: clustering and SVM learning.

The overall approach should be automatically run at very low frequency, such as every month, to capture possible modifications in water usage habits, both at urban and individual customer levels. At least one year should be selected as the observation period in order to identify possible cycles, seasonality and recurrent behaviours.

3. Results and Discussion

This section reports the results of the study, divided into the two different sources of data, urban water demand (SCADA) and individual customer consumption (AMR) data, and with respect to the different goals/services provided by the proposed approach, demand forecasting to support pump scheduling optimization, and anomaly detection.

3.1. SCADA Data and Urban Demand Forecasting

First, the results related to the first stage of the approach, that is, time-series data clustering, are reported. The two-level clustering approach was applied allowing for the identification of six typical daily urban water demand patterns, three seasons ($k_1 = 3$) and two different types of day for each season (${k_2}^i = 2$ with $i = 1, \ldots, k_1$), when $k_1 = 1, \ldots, 6$ and ${k_2}^i = 1, \ldots, 4$ were used.

Clustering in six clusters is the result that allows the best values of the Silhouette (0.74, averaged on the two-level clustering) and Calinski-Harabasz (97.87, averaged on the two-level clustering) indices. More in detail, the indices values for the other values of k_1 and k_2 are summarized in the following Tables 1 and 2.

Table 1. Silhouette and Calinski-Harabasz measures for different k_1 (i.e., the number of clusters at the first step of the proposed approach).

k_1	Silhouette	Calinski-Harabasz
2	0.51	2.96
3	0.74	3.15
4	0.49	2.12

Table 2. Silhouette and Calinski-Harabasz measures for different k_2 (i.e., the number of clusters at the second step of the proposed approach), where $k_1 = 3$ has been selected.

k_2	Silhouette	Calinski-Harabasz
2	0.70	192.78
3	0.62	54.35
4	0.50	36.68

The numeric result, related to the internal validation criteria adopted (Silhouette and Calinski-Harabasz), is also confirmed by the following "semantic" interpretation of the cluster assignment.

The following Figure 5 shows a calendar with the cluster assignment for every day of the observation period. This kind of visualization makes seasonality and cyclic behaviours more evident. The $k_1 = 3$ clusters at the first level can be identified as

- "Fall-Winter": from November to March
- "Spring-Summer": from April to June and from September to October
- "Summer break": July and August

The cluster "Fall-Winter" and "Spring-Summer" are further divided into "working days" and "holiday/weekends". The further clustering of the time-series data related to the cluster "Summer break" highlights the difference between the working days in July and the first and last week of August and the other days that are weekends, holidays or the summer vacation period in the middle of August. It is important to highlight that this calendar is an output of the clustering procedure; it is knowledge directly inferred from data, so any changes in consumption habits may be captured through the proposed approach.

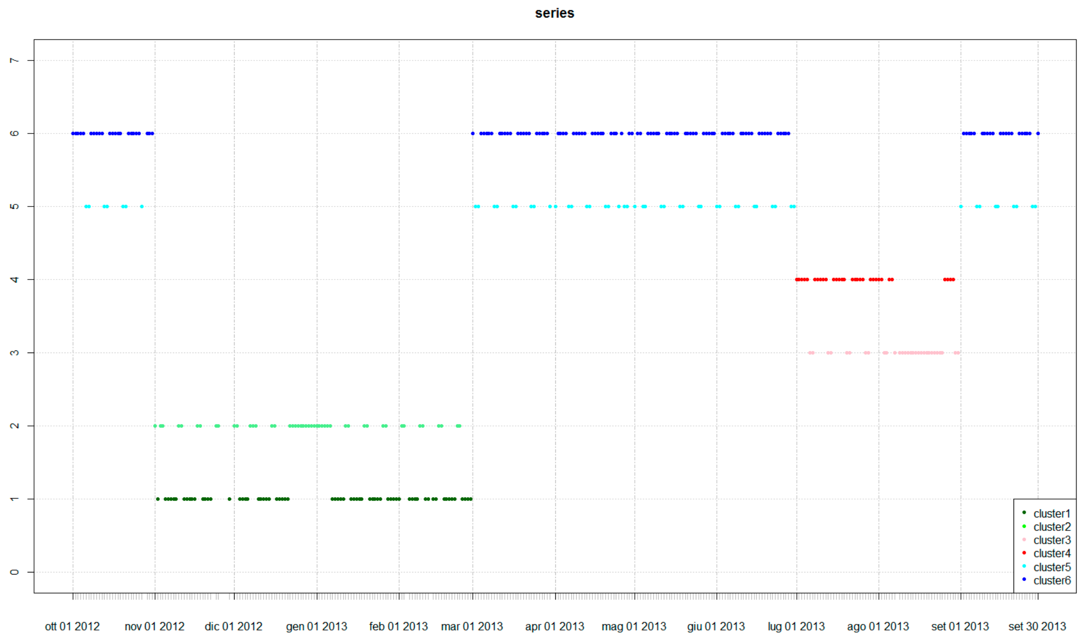


Figure 5. A graph showing the cluster assignment over the observation period. Although the final cluster label is used, it is easy to also identify the three clusters obtained at the first level; cluster1 & cluster2, cluster3 & cluster4, and cluster5 & cluster6.

In the following Figure 6, the centroids of the six different clusters are shown, representing the six different typical consumption behaviours. It is easy to note that major differences among the identified typical patterns highlight the peaks in consumption in the morning and in the evening. In particular, the peak in the morning of holidays and weekends is always delayed by approximately 1 h with respect to that of working days for each period of the year.

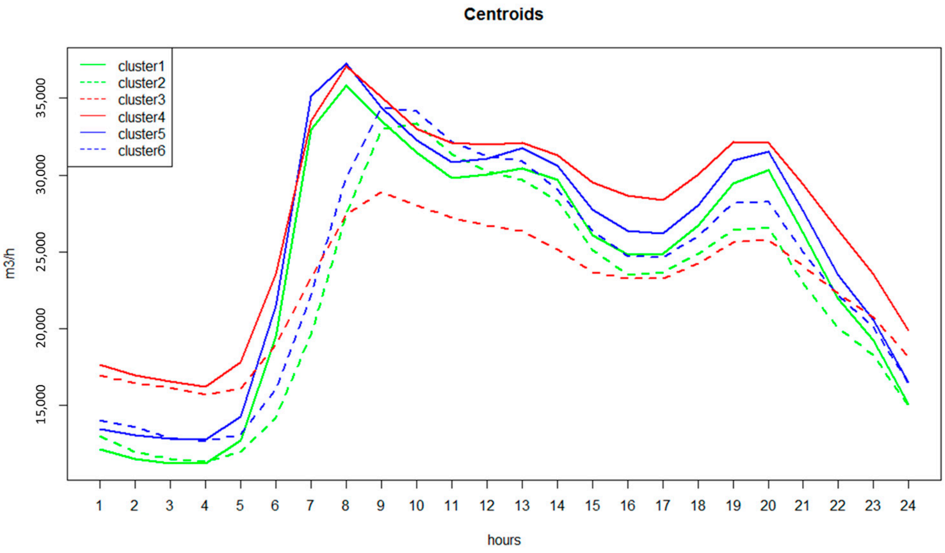


Figure 6. The *k* = 6 typical water demand patterns identified through the two-level clustering procedure. The cluster assignment is the same as the previous Figure 1.

After the clustering procedure, one SVM is trained for each cluster and for each hour of the day, by using the first six values of hourly consumption as input features (p = 6) and the hourly water consumption at one of the remaining hours of the day as the target variable.

Parameters of every SVM regression model are tuned independently by minimizing RMSE via leave-one-out-validation and taking into account several alternatives among linear, Polynomial and Radial Basis Function (RBF) kernels.

Since the overall forecasting for a day is performed through a pool of SVMs—one pool for each cluster is learned—the forecasting error is measured through MAPE. Table 3 summarizes the results obtained on each cluster that is the error provided by every SVM pool.

Table 3. Mean Absolute Percentage Error (MAPE) for the best and the worst forecasts in each cluster with standard deviation.

	Size	Best	Worst
Cluster 1	67	0.79% ± 0.59%	6.11% ± 2.95%
Cluster 2	46	1.57% ± 1.18%	14.33% ± 11.68%
Cluster 3	30	0.84% ± 0.66%	8.48% ± 3.53%
Cluster 4	31	1.71% ± 2.56%	12.84% ± 7.53%
Cluster 5	54	1.31% ± 0.93%	7.85% ± 13.26%
Cluster 6	127	1.10% ± 0.85%	6.54% ± 3.46%

3.2. AMR Data and Anomaly Detection

With respect to the AMR data, only two different clusters for each AMR were identified at the first stage of the approach. In this case, it was not possible to apply the two-level clustering procedure due to the small observation period that does not allow for the identification of seasonality but just types of days. This is a common result, both for domestic and non-domestic customers; daily water consumption patterns are logically different according to the different types of water usage.

As representative results, some typical patterns identified for three different AMRs are reported in the following Figures 7–9: a residential, a non-residential, and a “mixed” customer, where “mixed” means that both residential and bot residential users are in the same building monitored by just one smart meter.

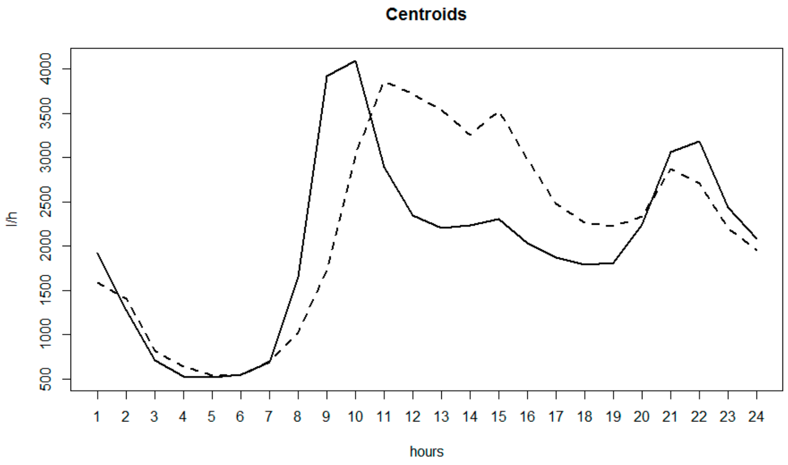


Figure 7. Case 1—a residential customer: results from clustering (dotted line is the pattern associated with weekends and holidays).

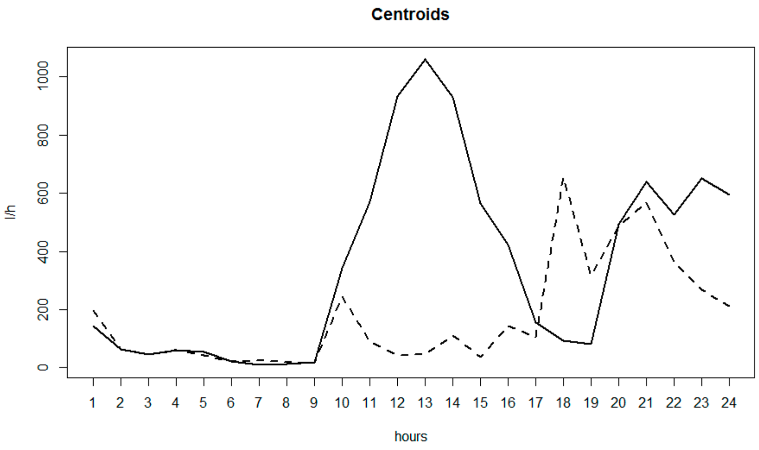


Figure 8. Case 2—a non-residential customer: results from clustering (dotted line is the pattern associated with Sunday, Monday, and holidays).

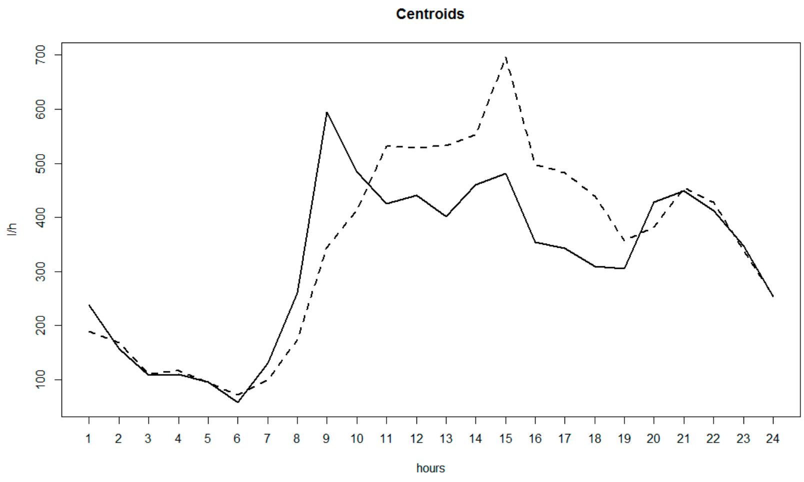


Figure 9. Case 3—a non-residential "mixed" customer: results from clustering (dotted line is the pattern associated with weekends and holidays).

In the second stage of the approach, a series of SVM regression models was learned, one SVM for each AMR, for each cluster, and for each hour. The same experimental setting used for SCADA data was applied in the AMR case; the water consumption values of the first six hours of the day ($p = 6$) were considered as input variables of an SVM regression model, while the target variable is a specific hourly consumption prediction. The best parameter setting for every SVM model was identified through leave-one-out validation and among a series of possible configurations of linear, Polynomial and Radial Basis Function kernels.

Although the overall approach is the same applied to the SCADA data, the forecasting has a completely different goal. While urban demand forecasting in the short term can effectively improve the effectiveness and efficiency of operations in the water networks, particularly the optimization of the pumping schedule, predicting the individual water consumption for the 24 h may improve customer relationship management through more regular and targeted demand-side management strategies, simulate and predict/prevent critical situations in the network by increasing resilience at structural and service level, and identify anomalies such as malfunctions, frauds or cyber-physical attacks (e.g., through False Data Injection).

The main aim of the validation of the proposed approach on the available set of AMR data is related to anomaly detection, defined as a significant deviation of the actual consumption pattern with respect to the forecast. Deviations from predicted typical consumption behaviour may occur because of meter faults, water theft through bypass or the transmission of false data. In this case, MAPE can be used to evaluate the entity of a deviation and identify possible anomalies, and it is not a simple prediction error measure.

Logically, MAPE computed on AMR data is higher than the one computed on SCADA data. This is mainly due to higher variability in the behaviour at the individual customer level and to the limited observation period (three months), which does not allow for dealing with uncertainty effectively. However, the variability of MAPE may be very different depending on the type of customer; in particular, variability was higher for non-residential customers, who are less characterized by recurrent/typical behaviours.

The following Figure 10 shows the value of MAPE over the observation period for the three different AMRs reported as representative cases.

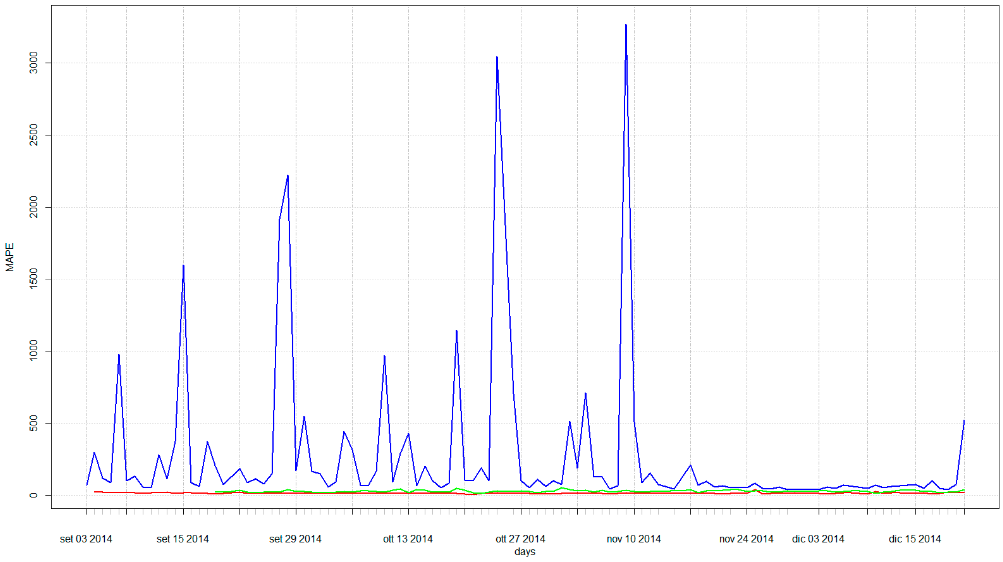


Figure 10. Mean Absolute Percentage Error (MAPE) over the observation period for the three cases (red, blue and green are residential, non-residential and non-residential mixed customers, respectively).

In the residential case, MAPE does not show a very high variability (as shown in the following Figure 11), and only a limited set of actual daily patterns can be considered "anomalous" with respect to the forecasts.

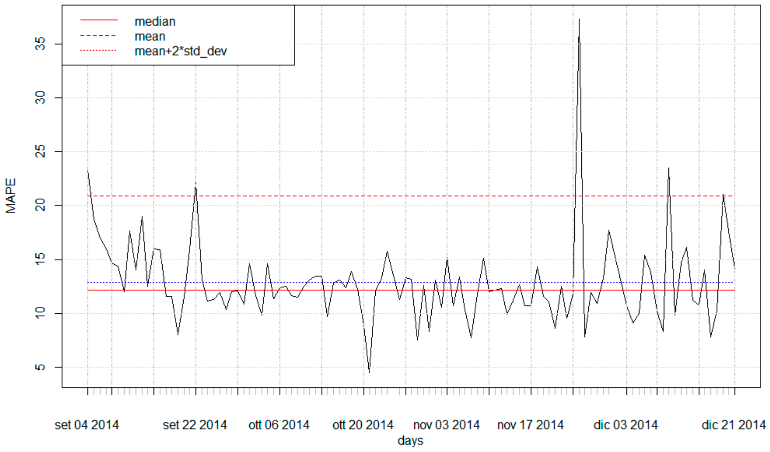


Figure 11. MAPE over the observation period for the residential case (Case 1).

The following Figure 12 compares the actual consumption pattern with the forecast pattern corresponding to the highest MAPE value.

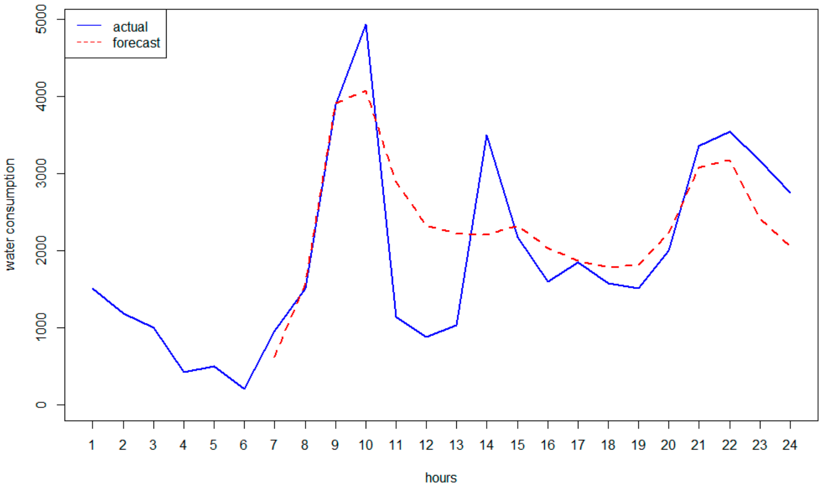


Figure 12. Case 1—actual consumption versus forecast for the day with the highest MAPE.

In the non-residential case, MAPE has very high variability (as shown in the following Figure 13), and several actual patterns are considered "anomalous" when compared to the associated forecasts.

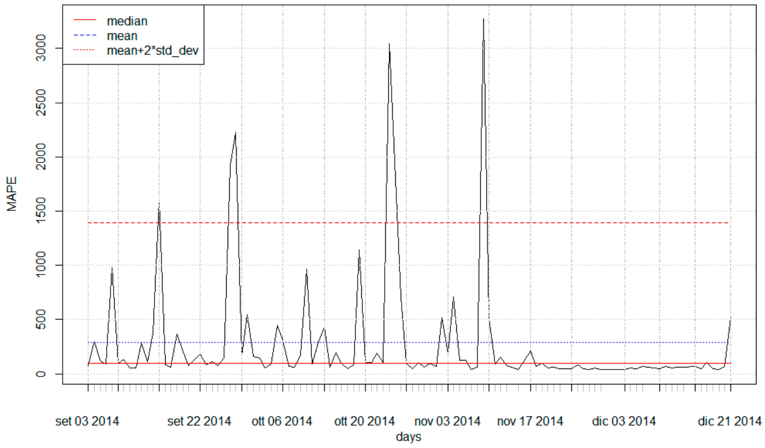


Figure 13. MAPE over the observation period for the non-residential case (Case 2).

In the following Figure 14, it is possible to understand when two different situations may occur: the variability in the consumption habits reduces the chance to obtain precise forecasts—but the overall forecasted pattern is valid—and a completely unexpected behaviour occurs. In the (b) side of Figure 10, an actual consumption value near zero is measured against a predicted consumption. This is an important anomaly to consider for generating a warning; this specific situation, for instance, may be associated with a smart meter fault or a possible fraud.

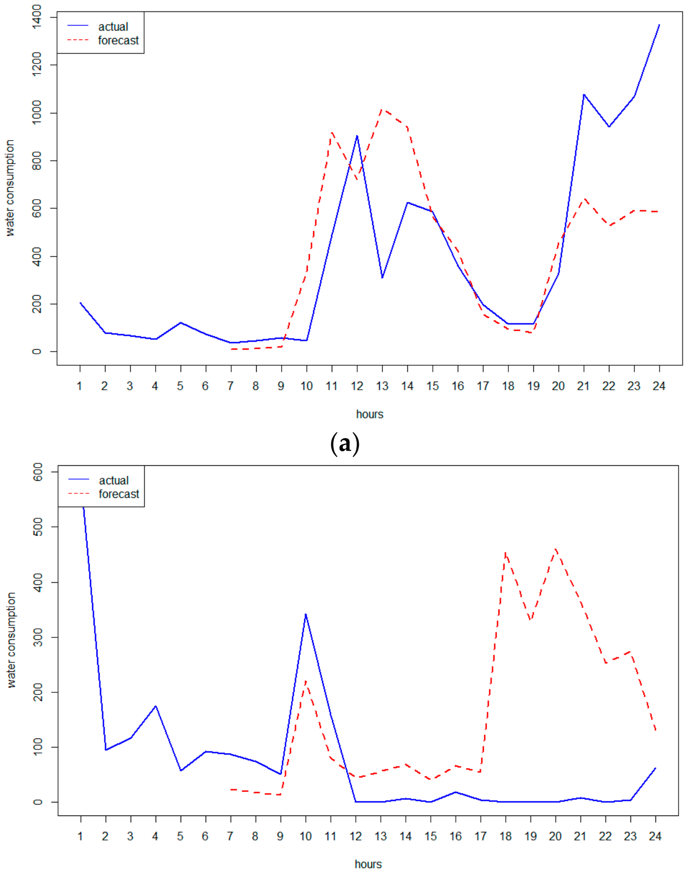


Figure 14. Case 2—actual consumption versus forecast for the day with the lowest (**a**) and highest (**b**) MAPE.

Finally, the following Figure 15 shows the variability of the MAPE from the non-residential "mixed" case. Analogous to the residential case, a limited number of actual patterns can be considered anomalous when compared to the relative forecasts.

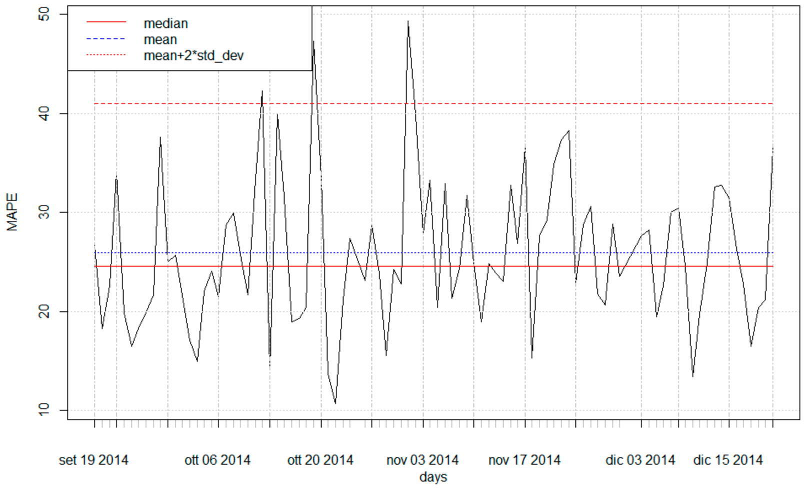
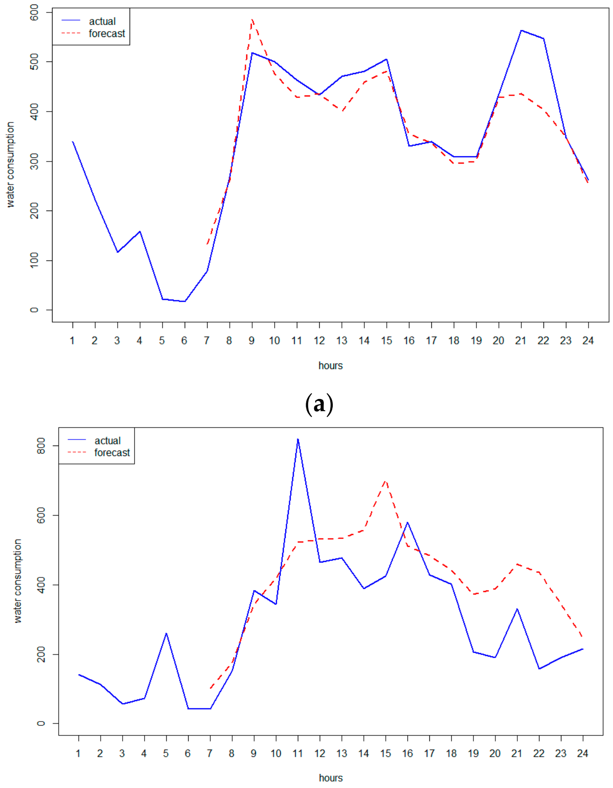


Figure 15. MAPE over the observation period for the non-residential "mixed" case (Case 3).

The comparison between actual and forecasted patterns, with respect to the days with lowest and highest MAPE, is reported in the following Figure 16. In this case, the value of MAPE does not identify any anomalous behaviour but only some forecasting errors or, at the most, a slight difference in the current consumption with respect to the usual habits.



(**a**)

(**b**)

Figure 16. Case 3—actual consumption versus forecast for the day with the lowest (**a**) andhighest (**b**) MAPE.

4. Conclusions

This paper presents a two-stage approach, based on time-series clustering and SVM regression, for analysing time-series data of water demand both at an aggregated level (urban water demand—SCADA data) and an individual level (single customer—AMR data).

With respect to the current state of the art in water demand forecasting, the proposed approach offers a number of innovations. First, it improves forecasting accuracy by exploiting the idea that usually just a limited set of typical water usage behaviours occurs, both at the urban and individual customer level; they can be easily inferred directly from raw data through a suitable time-series clustering procedure (first stage of the approach). These behaviours may recur on different time scales, such as seasonality, a monthly time scale, and the type of day, at a finer scale, and the corresponding representative patterns—centroids of the clusters—are different in terms of the occurrence of peaks and bursts at specific hours of the day and linked to water usage habits. Second, various SVM regression models are learned, in particular one for each cluster and for each hour's consumption, in order to increase forecasting accuracy by exploiting the benefits provided by this specific regression method and currently are widely adopted in the literature (but usually by learning a unique SVM regression model).

Most important, the proposed approach proved to be applicable—as is—on both urban water demand data and individual customer consumption data. The two different applications allow for addressing two specific issues in urban water distribution network management, that is, urban demand forecasting and anomaly detection. While the first is devoted to supporting the optimization of operations, in particular the optimization of the pumping schedule, the second identifies online possible anomalies due to smart meter faults, frauds, cyber-physical attacks, and significant shifts in water usage habits.

The error measure used in the study (MAPE) is the most widely adopted in the water demand forecasting literature and enables a comparison of this approach to other studies and approaches. However, results on AMR data show that a different interpretation of the MAPE values is required; the variability of individual water usage habits is higher, in particular for non-residential customers, and high values in MAPE could be anomalies instead of simply incorrect predictions.

Future work will address scalability issues in the case of a massive number of AMRs and the possibility to implement an online/stream learning version of the proposed approach. Finally, a variation of the MAPE—or a suitable analysis of its values—will be proposed in the case of forecasting individual consumption data in order to better differentiate between forecasting errors and occurrences of anomalies.

Acknowledgments: This work was partially performed within ICeWater, which received funding from the European Union's Seventh Framework Programme for research, technological development and demonstrations under Grant Agreement No. 317624.

Author Contributions: Antonio Candelieri conceived and designed the analytical approach proposed; performed analysis, and wrote the paper.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Mamo, T.G.; Juran, I.; Shahrour, I. Urban water demand forecasting using the stochastic nature of short term historical water demand and supply pattern. *J. Water Resour. Hydraul. Eng.* **2013**, *2*, 92–103.
2. Bakker, M.; van Duist, H.; van Schagen, K.; Vreeburg, J.; Rietveld, L. Improving the performance of water demand forecasting models by using weather input. *Proced. Eng.* **2014**, *70*, 93–102. [CrossRef]
3. New York City Environment Protection. Available online: http://www.nyc.gov/html/dep/html/press_releases/10-78pr.shtml#.WHUdDVPhCUk (accessed on 16 January 2017).
4. 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. *J. Water Supply: Res. Technol. AQUA* **2013**, *62*, 1–13. [CrossRef]

5. Sebri, M. Forecasting urban water demand: A meta-regression analysis. *J. Environ. Manag.* **2016**, *183*, 777–785. [CrossRef] [PubMed]
6. Donkor, E.A.; Mazzucchi, T.A.; Soyer, R.; Roberson, J.A. Urban water demand forecasting: review of methods and models. *J. Water Resour. Plan. Manag.* **2014**, *140*, 146–159. [CrossRef]
7. Romano, M.; Kapelan, Z. Adaptive water demand forecasting for near real-time management of smart water distribution systems. *Environ. Model. Softw.* **2014**, *60*, 265–276. [CrossRef]
8. Know, H.; So, B.; Kim, S.; Kim, B. Development of ensemble model based water demand forecasting model. *EGU Gen. Assem. Conf. Abstr.* **2014**, *16*, 3711.
9. Gargano, R.; Tricarico, C.; Del Giudice, G.; Granata, F. A stochastic model for daily residential water demand. *Water Science and Technology: Water Supply* **2016**, *16*, 1753–1767. [CrossRef]
10. Magini, R.; Pallavicini, I.; Guercio, R. Spatial and temporal scaling properties of water demand. *J. Water Resour. Plann. Manage.* **2008**, *134*, 276–284. [CrossRef]
11. Alcocer-Yamanaka, V.H.; Tzatchkov, V.G.; Arreguin-Cortes, F.I. Modeling of drinking water distribution networks using stochastic demand. *Water Resour. Manage.* **2012**, *26*, 1779–1792. [CrossRef]
12. Blokker, E.J.M.; Vreeburg, J.H.G.; van Dijk, J.C. Simulating residential water demand with a stochastic end-use model. *J. Water Resour. Plann. Manage.* **2010**, *136*, 19–26. [CrossRef]
13. Buchberger, S.G.; Wu, L. A model for instantaneous residential water demands. *J. Hydraul. Eng.* **1995**, *121*, 232–246. [CrossRef]
14. Gargano, R.; Di Palma, F.; de Marinis, G.; Granata, F.; Greco, R. A stochastic approach for the water demand of residential end users. *Urban. Water J.* **2016**, *13*, 569–582. [CrossRef]
15. Granata, F.; Papirio, S.; Esposito, G.; Gargano, R.; de Marinis, G. Machine Learning Algorithms for the Forecasting of Wastewater Quality Indicators. *Water* **2017**, *9*, 105. [CrossRef]
16. Wu, M.C.; Lin, G.F. An Hourly Streamflow Forecasting Model Coupled with an Enforced Learning Strategy. *Water* **2015**, *7*, 5876–5895. [CrossRef]
17. Adamowski, J.; Karapataki, C. Comparison of multivariate regression and artificial neural networks for peak urban water-demand forecasting: evaluation of different ann learning algorithms. *J. Hydrol. Eng.* **2010**, *15*, 729–743. [CrossRef]
18. Cutore, P.; Campisano, A.; Kapelan, Z.; Modica, C.; Savic, D. Probabilistic prediction of urban water consumption using the scem-ua algorithm. *Urb. Water J.* **2008**, *5*, 125–132. [CrossRef]
19. Firat, M.; Yurdusev, M.A.; Turan, M.E. Evaluation of artificial neural network techniques for municipal water consumption modeling. *Water Resour. Manag.* **2009**, *23*, 617–632. [CrossRef]
20. Herrera, M.; Torgo, L.; Izquierdo, J.; Pérez-García, R. Predictive models for forecasting hourly urban water demand. *J. Hydrol. Eng.* **2010**, *387*, 141–150. [CrossRef]
21. Ghiassi, M.; Zimbra, D.; Saidane, H. Urban water demand forecasting with a dynamic artificial neural network model. *J. Water Resour. Plan. Manag.* **2008**, *134*, 138–146. [CrossRef]
22. Herrera, M.; Izquierdo, J.; Pérez-Garćıa, R.; Ayala-Cabrera, D. On-line learning of predictive kernel models for urban water demand in a smart city. *Proced. Eng.* **2014**, *70*, 791–799. [CrossRef]
23. Ji, G.; Wang, J.; Ge, Y.; Liu, H. Urban Water Demand Forecasting by LS-SVM with Tuning Based on Elitist Teaching-Learning-Based Optimization. In Proceedings of the 26th Chinese Control and Decision Conference (2014 CCDC), Changsha, China, 31 May–2 June 2014; pp. 3997–4002.
24. Sampathirao, A.K.; Grosso, J.M.; Sopasakis, P.; Ocampo-Martinez, C.; Bemporad, A.; Puig, V. Water Demand Forecasting for the Optimal Operation of Large-Scale Drinking Water Networks: The Barcelona Case Study. In Proceedings of the 19th International Federation of Automatic Control (IFAC) World Congress, Cape Town, South Africa, 2014; pp. 10457–10462.
25. Brentan, B.; Luvizotto, E.; Herrera, M.; Izquierdo, J.; Perez-Garcia, R. Real-time water demand forecasting using support vector machine and adaptive fourier series. In *Modelling for Engineering and Human Behaviour*; Jodar, L., Acedo, L., Cortes, J.C., Eds.; IMM-Universitat Politecnica de Valencia: Valencia, Spain, 2015; pp. 178–182.
26. Bai, Y.; Wang, P.; Li, C.; Xie, J.; Wang, Y. Dynamic forecast of daily urban water consumption using a variable-structure support vector regression model. *J. Water Resour. Plan. Manag.* **2015**, *141*, 04014058. [CrossRef]
27. Rao, R.V.; Savsani, V.J.; Vakharia, D.P. Teaching–learning-based optimization: a novel method for constrained mechanical design optimization problems. *Comput. Aided Des.* **2011**, *43*, 303–315. [CrossRef]

28. Shabani, S.; Yousefi, P.; Adamowski, J.; Naser, G. Intelligent soft computing models in water demand forecasting. In *Water Stress in Plants*; Rahman, M.I.M., Begum, Z.A., Hasegawa, H., Eds.; Intech: Rijeka, Croatia, 2016; Chapter 6, pp. 100–117.
29. Martìnez-Alvarez, F.; Troncoso, A.; Riquelme, J.C.; Aguilar-Ruiz, J.S. Energy time series forecasting based on pattern sequence similarity. *IEEE Trans. Knowl. Data Eng.* **2011**, *23*, 1230–1243. [CrossRef]
30. Bokde, N.; Asencio-Cortés, G.; Martínez-Álvarez, F.; Kulat, K. Psf: Introduction to r package for pattern sequence based forecasting algorithm. arXiv:1606.05492v2 [stat.ML] 25 Aug 2016.
31. Candelieri, A.; Soldi, D.; Archetti, F. Short-term forecasting of hourly water consumption by using automatic metering readers data. *Proced. Eng.* **2015**, *119*, 844–853. [CrossRef]
32. Candelieri, A.; Archetti, F. Identifying typical urban water demand patterns for a reliable short-term forecasting – the icewater project approach. *Proced. Eng.* **2014**, *89*, 1004–1012. [CrossRef]
33. Liao, T.W. Clustering of time series data—a survey. *Pattern Recognit.* **2005**, *38*, 1857–1874. [CrossRef]
34. Kavitha, V.; Punithavalli, M. Clustering time series data stream-a literature survey. *J. Comput. Sci. Inf. Secur.* **2010**, *8*.
35. Zhang, X.; Liu, J.; Du, Y.; Lv, T. A novel clustering method on time series data. *Expert Syst. Appl.* **2011**, *38*, 11891–11900. [CrossRef]
36. Maitra, R.; Ramler, I.P. A k-mean-directions algorithm for fast clustering of data on the sphere. *J. Comput. Gr. Stat.* **2010**, *19*, 377–396. [CrossRef]
37. Arbelaitz, O.; Gurrutxaga, I.; Muguerza, J.; Pérez, J.M.; Perona, I. An extensive comparative study of cluster validity indices. *Pattern Recognit.* **2013**, *46*, 243–256. [CrossRef]
38. Vapnik, V.N. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 1995.
39. Vapnik, V.N. *Statistical Learning Theory*; Wiley: New York, NY, USA, 1998.
40. Schölkopf, B.; Smola, A.J. *Learning with Kernels: Support. Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, USA, 2002.