A Mixed Strategy Based on Self-Organizing Map for Water Demand Pattern Profiling of Large-Size Smart Water Grid Data

Roberta Padulano[1] · Giuseppe Del Giudice[1]

Received: 31 January 2018 / Accepted: 14 May 2018

Abstract In the present paper a procedure is introduced to detect water consumption patterns within water distribution systems. The analysis is based on hourly consumption data referred to single-household flow meters, connected to the Smart Water Network of Soccavo (Naples, Italy). The procedure is structured in two consecutive phases, namely clustering and classification. Clustering is performed on a selection of standardized monthly time series, randomly chosen within the database; different clustering models are tested, basing on K-means, dendrogram and Self-Organizing Map, and the most performant is identified comparing a selection of Clustering Validity Indices. Supervised classification is performed on the remaining time series to associate unlabeled patterns to the previously defined clusters. Final results show that the proposed procedure is able to detect annual patterns describing significant customers behaviors, along with patterns related to instrumental errors and to abnormal consumptions.

Keywords Calinske–Harabasz index · Classification · Clustering · Davies–Bouldin Index · Dendrogram · K-means · Mixed strategy · Self-Organizing Map · Water demand modeling · Water demand patterns

1 Introduction

Water demand modeling and forecast is a key issue in modern approaches to an efficient water management. A comprehensive knowledge of water consumption allows for a correct

✉ Roberta Padulano
roberta.padulano@unina.it

Giuseppe Del Giudice
giuseppe.delgiudice@unina.it

[1] Department of Civil, Architectural and Environmental Engineering, Università degli Studi di Napoli Federico II, Naples, Italy

Published online: 26 May 2018

planning of water supply, for the estimate of leakages in the water distribution networks and for the development of innovative approaches and attractive plans to consumers. The increasing interest toward water systems efficiency has led to the implementation of "Smart Water Grids" within urban areas, with significant portions of customers connected to a telemetry system for flow data reading and collection. Smart grids allow for the collection of large amounts of data, usually on an hourly basis or less (Gargano et al. 2016; Cominola et al. 2018) which can be used to understand consumption drivers at the customer scale. This is a challenging task in a complex urban environment because of the extreme variability in the characteristics of households, such as the number of individuals served by each flow meter, water usage (which can be related to either residential or commercial activities), different life habits of the end users.

For water demand modeling several methodologies can be applied, depending on the specific goal. One common approach consists in treating water consumption as stochastic data and looking for probability distribution models for different times of the day (Gargano et al. 2016). Stochastic models are also used for instantaneous residential water demand requiring parameters such as the frequency, duration and intensity of a single consumption event which are quite difficult to achieve (Blokker et al. 2010; Fontanazza et al. 2016). Other models exist that aim at explaining the deterministic components of water demand, relating significant consumption indices with socio-demographic, billing and climate factors (Arbués et al. 2010; Browne et al. 2013; Parker and Wilby 2013; Loureiro et al. 2016; Bergel et al. 2017; Ghavidelfar et al. 2017; Haque et al. 2017). One last approach consists in the profiling, namely a detection of demand patterns based on a large amount of consumption data; this is a typical approach in the electricity sector (Räsänen et al. 2010; López et al. 2011; Ferreira et al. 2013; Zhou et al. 2013; Macedo et al. 2015), with a few applications for water demand modeling (McKenna et al. 2014; Avni et al. 2015). Profiling of consumption data is typically performed to catch differences in the customers behavior, with particular focus on the weekdays/weekends distinction, especially when no previous information is known.

The present paper provides a methodology aiming at analyzing hourly water consumption data recorded within a very large Smart Water Grid in order to profile annual patterns detecting significant behaviors and life habits of residential customers. As main novelty element, pattern detection is performed by means of a clustering/classification procedure based on Self-Organizing Map, whose applications in the water demand framework are very limited. In the following sections, a theoretical background will be given concerning clustering and classification methods, along with the description of the District Metering Area.

2 Materials and Methods

2.1 Clustering

Clustering is a data mining technique that consists in dividing an initial set of multidimensional data in different meaningful subsets containing objects that share similar characteristics, with the aim of discovering hidden recurring patterns (Zhou et al. 2013; Sancho-Asensio et al. 2014). An efficient clustering provides a number of final clusters such that the distance among data belonging to different clusters (usually referred to as "between-clusters distance") is maximized, whereas the distance among data belonging to the same clusters ("within-clusters distance") is minimized (López et al. 2011; Avni et al. 2015). Given a couple of multidimensional data $\vec{x}_i$ and $\vec{x}_j$, the most common definition of their reciprocal distance $d(\vec{x}_i, \vec{x}_j)$ is the Euclidean norm (Keogh et al. 2001; Popivanov and Miller 2002):

$$d(\vec{x}_i, \vec{x}_j) = \sqrt{(\vec{x}_i - \vec{x}_j) \cdot (\vec{x}_i - \vec{x}_j)^T} \qquad (1)$$

A common classification of clustering techniques is among “partitioning” (i.e. K-means), “hierarchical” (i.e. dendrogram) and “model-based” methods (i.e. Self-Organizing Maps) (Zhou et al. 2013). K-means is a clustering algorithm consisting in the crisp partition of multidimensional data into K subsets, with the number of clusters K defined by the user prior to the analysis (MacQueen et al. 1967). The partition is made assigning each data to the nearest cluster center, or “centroid”; initial cluster centroids are assigned randomly and any available distance metric can be used. A number of iteration must be run in order to minimize the effect of the initial cluster centroids choice (Zhou et al. 2013). The dendrogram method consists in a bottom-up agglomeration of data based on reciprocal distances (Johnson 1967). Starting from a condition where each data is a separate cluster, pairwise distances are computed and the two nearest data are merged into a new cluster, whose centroid is evaluated and pairwise distances are updated. The merging of couples of clusters continues until the desired number of clusters, which is defined by the user prior to the analysis, is achieved (Jota et al. 2011; Zhou et al. 2013). Self-Organizing Map (SOM) is an unsupervised clustering technique based on neural networks (Kohonen 1982). The main concept is that an input layer made up of initial data must be reduced in size and connected to an output layer by means of network parameters and adjustable weights (Kalteh et al. 2008). The output layer usually consists of a bidimensional grid made up of a number K of typically hexagonal elements, or “output neurons”, which represent the maximum possible number of clusters, defined by the user prior to the analysis. The final output of SOM is represented by the output neuron grid, where each neuron can be empty, if no input data was found to be related to it, or filled with one or more input data, so that each neuron can be regarded to as a separate cluster. However, the in-deep insight of SOM results, implying visual inspection of neighboring distances and component planes, and also taking into account additional information such as data labels (Verdú et al. 2006; Laspidou et al. 2015), can lead to a merging of the nearest neurons with the consequent reduction of clusters.

There is no confirmation in literature about which is the most performant clustering algorithm, since each method has both advantages and drawbacks (Räsänen et al. 2010; López et al. 2011; Zhou et al. 2013). However, different methods are usually coupled with specific applications. As concerns smart metering data, several examples exist that adopt SOM clustering for energy consumption data to recognize multiple consumers typologies or to discriminate weekdays from weekends consumption (Verdú et al. 2006; Räsänen et al. 2010; Macedo et al. 2015). In the field of water consumption pattern analysis, recent applications include K-means (Avni et al. 2015) and SOM (Laspidou et al. 2015). Dendrogram applications are rare, since this method implies a deep computational effort to compute the dissimilarity matrix when the initial dataset is large (Schikuta 1996).

Whichever the algorithm used for clustering, one key parameter is the number of clusters K (or its equivalent for SOM, namely the output grid dimension). In practical applications, some prior information is usually available so that the choice of K is data-driven and not arbitrary (for instance, when using K-means the number of consumers typologies should be known). As concerns SOM, the choice of the output grid dimension (which is the square root of the maximum allowed number of neurons) can follow two different strategies. The first approach consists in setting a very large output dimension in order to obtain an output map made up of groups of neurons occupied by one or few data, delimited by groups of empty neurons. This method is usually applied when there is a prior knowledge about data labeling (for example the corresponding day of the week is known) and the merging of close neurons into clusters is straightforward. At the opposite, when no prior information or labeling is available, it is more useful to set a small grid dimension and let each neuron be hit by a considerable number of data. When this happens, the output neurons coincide with the final clusters, although a posterior merging of the nearest neurons can always be considered.

In general, when prior knowledge is little or absent, it is a common practice to repeat computations with different values of K and compare results, looking for the best "cluster solution" in terms of clustering quality. The performance of a cluster solution can be evaluated by means of the Clustering Validity Indices (CVIs). A large number of CVIs has been proposed in literature (Dimitriadou et al. 2002; Räsänen et al. 2010) and there is no general consensus about which should be the most useful (Zhou et al. 2013). A general approach is to pick the cluster solution that either minimizes/maximizes a certain CVI or corresponds to an elbow/local peak of the function (Dimitriadou et al. 2002). However, it is important to understand that taking more than one CVI into account at the same time could lead to problematic clustering evaluation, because multiple CVIs seldom give the same results; the choice of which CVI to use and the interpretation of results should be considered heuristic (Dimitriadou et al. 2002).

Among all the proposed CVIs the most frequently used are based on the definition of between-clusters distance SSB and within-clusters distance SSW, which account for the definition of distance proposed in Eq. 1:

$$SSB = \sum_{k=1}^{K} n_k \cdot d^2(\vec{C}_k, \vec{M}) \tag{2}$$

$$SSW = \sum_{k=1}^{K} \sum_{i=1}^{n_k} d^2(\vec{x}_i, \vec{C}_k) \tag{3}$$

where n_k is the number of data in cluster k, $\vec{C}_k$ is the centroid of cluster k and $\vec{M}$ is the mean of all data in the dataset. Whichever the algorithm adopted and the initial dataset, by definition SSW increases and SSB decreases for increasing K, and their sum remains constant (López et al. 2011). SSB and SSW can be used in combination with other CVIs or even alone to make some preliminary cluster evaluation: for example, the best cluster number K could be chosen as that value where SSB and SSW stabilize to an asymptote (Räsänen et al. 2010). Basing on Eqs. 2 and 3, two CVIs were proposed that are the most frequently used for clustering evaluation, namely the Calinski-Harabasz index CH (Caliński and Harabasz 1974) and the Davies-Bouldin Index DBI (Davies and Bouldin 1979):

$$CH = \frac{SSB}{SSW} \cdot \frac{N-K}{K-1} \tag{4}$$

where N is the number of data in the initial dataset, and

$$DBI = \frac{1}{K} \cdot \sum_{k=1}^{K} max(R_{kj}) \tag{5}$$

with:

$$R_{kj} = \frac{\bar{d}_k + \bar{d}_j}{d(\vec{C}_k, \vec{C}_j)} \quad \forall k, j \in K \tag{6a}$$

$$\bar{d}_k = \frac{1}{n_k} \cdot \sum_{i=1}^{n_k} d(\vec{x}_i, \vec{C}_k) \tag{6b}$$

Computation of CH is straightforward once SSB and SSW have been calculated, whereas for the computation of DBI some successive steps must be accomplished. First, for each cluster k in the cluster solution the mean $\bar{d}_k$ ("average within distance") of all distances of data in the cluster from the cluster centroid must be computed. Then, for each pair of clusters in the cluster solution the quantity R_{kj} must be computed which is the sum of average within distances of the clusters in the pair, normalized by the distance between the two centers. Finally, for each cluster k the maximum of R_{kj} is found, and DBI is the average of maximum R_{kj} values. The best cluster solution is the one that minimizes DBI or maximizes CH.

As seen, the state of the art provides for different techniques, along with different approaches to set relevant parameters and performance criteria, and there is no algorithm nor CVI performing suitably well in every context. For this reason, in the present paper a "mixed clustering strategy" is adopted that consists in combining different techniques to detect clusters in very large datasets, where single methods could perform poorly (Lebart et al. 2004; Bocci and Mingo 2012). Specifically, a base partitioning will be obtained with a first-level clustering by SOM, and a second-level clustering will be performed to the centroids of the first-level clusters. For first-level clustering different output grid dimension values will be tested, whereas for second-level clustering both K-means and dendrogram will be applied with different cluster numbers K. Clustering performances will be compared by means of DBI and CH indices.

2.2 Classification

Classification is a supervised clustering technique aiming at grouping data in different subsets on the basis of a pre-existing labeling. Classification usually consists in two steps: in the first step, each input data is labeled as belonging to a specific cluster, and a training occurs to build a model that is able to recognize which are the intrinsic characteristics of each cluster. In the second step, the so trained model is able to associate novel unlabeled data to each of the modeled clusters (Zhu 2006; Papa et al. 2009).

Different classifiers are available for the training step, such as decision trees (Breiman et al. 1984), nearest neighbor (Friedman et al. 1977), discriminant analysis (Krzanowski 1988) and Support Vector Machines (Cristianini and Shawe-Taylor 2000). A common procedure is to test different training models and choose the most performant one. To do this, several performance indicators can be used that give similar information arranged in different ways, such as the confusion matrix and the Receiver Operating Characteristics (ROC) curves. The confusion matrix enables the visualization of the training performance: each row represents the instances in a defined cluster while each column represents the instances in the predicted cluster (Fawcett 2006; Powers 2007). The percentages in the diagonal must be as high as possible since they represent the amount of data belonging to each cluster that were classified as belonging to that cluster ("True Positive Rate" TPR). If one of the elements in the diagonal is not equal to 100%, the remaining data will be found on the same row but in a different column, since they will have been wrongly labeled ("False Negative Rate" FNR).

A ROC curve, one for each cluster, shows the False Positive Rate FPR (namely the amount of data labeled as belonging to a specific cluster but actually belonging to one of the remaining clusters) as a function of the true positive rate TPR with varying model parameter (Fawcett 2004; Briggs and Zaretzki 2008). The main elements of a ROC curve are the Current Classifier Point, whose coordinates are (0,1) if the model correctly recognizes all the data in that cluster, and the Area Under the Curve (AUC) that is equal to 1 in the same condition.

2.3 District Metering Area

The District Metering Area (DMA) which is the subject of the study is located in the North-Western part of the City of Naples (Italy) (Fig. 1). This area was chosen as a pilot area for a smart Water Distribution Network (WDN) implementation, with particular focus on the remote monitoring of flow meters, as part of a cooperation between the University of Naples and ABC – Napoli, which is the Municipal water company. The DMA is provided with 4254 customer connections whose flow meters were completely replaced during the last three years. The connections consist of 3701 (87% of the total number) residential flow meters, whereas the remaining 553 (13%) corresponds to commercial flow meters, consistently with the residential purpose of the neighborhood. Moreover, 2999 (81% of the residential flow meters) connections relates to single households, whereas the remaining 702 (19%) flow meters serve multiple households, such as duplexes or whole apartment buildings.

The present paper focuses on single-household flow meters; for each flow meter, 12 months of hourly consumption measurements are available, dating 01/01/2016 to 31/12/2016. A preliminary analysis on the time series was performed to detect (and replace, if needed) outliers in accordance with the MAD criterion (Cousineau and Chartier 2010); this also caused the elimination of 18% time series which were considered unreliable due to their large number of outliers. Therefore, the final database consists in 2454 time series of hourly data covering the period 01/01/2016-31/12/2016. Each time series is made up of a maximum of 8784 hourly data, 366 daily data and 12 monthly data. From the database 168 time series were randomly extracted to calibrate possible clusters ("clustering sample"); the remaining 2286 constitute the "classification sample" for clustering validation.

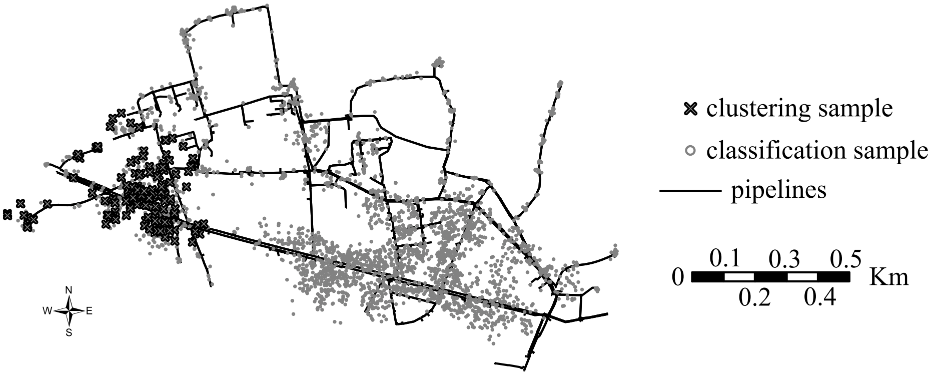


Fig. 1 Soccavo Smart Water Network: pipelines and monitored flow meters

3 Discussion of Results

Clustering was performed on the monthly aggregated time series belonging to the clustering sample in order to look for typical annual consumption patterns; the monthly scale was preferred to daily aggregation since it is less sensitive to anomalous daily behavior of consumers. Monthly time series were standardized by dividing each "monthly discharge" Q_m (namely the monthly mean of hourly volumes, computed for each month and for each flow meter) by the related annual mean discharge Q_a. Standardization was required to compare patterns of time series having different Q_a values.

No preliminary information is available that suggests a possible reliable cluster number K, nor literature points toward a particular clustering technique. As a consequence, different methods were applied and results were compared to find the optimal cluster solution. The in-deep analysis of cluster solutions required the estimation of several CVIs (CH and DBI were chosen) which, as stated in the previous sections, must be considered as a heuristic decision tool to be supported by additional observations about clusters consistency and meaningfulness.

Table 1 shows the different models applied to perform clustering, along with some significant parameters. Model A and model B consist in the application of two ordinary techniques, namely K-means and dendrogram respectively, to the initial set made up of 168 monthly-aggregated normalized time series. For both models the algorithm was run with 27 different cluster numbers (henceforth called K_2) chosen in the range 2–64. Models C consist in a mixed strategy made up of a first-level clustering by SOM (with output grid dimension identified by the first number following letter "C") and a second-level clustering by K-means (second number following letter "C" is 1) or dendrogram (second number following letter "C" is 2). For models C, SOM was run with 5 different output grid dimension $\sqrt{K_1}$ ranging between 4 and 8 (Table 1), where K_1 is the number of clusters of the 1st-level clustering. However, it must be noted that for high K_1 values not all the neurons could be occupied (for instance in models C81 and C82 the largest cluster solution is $K_1 = 63$ because one neuron was found empty), so that K_1 should be interpreted as the maximum

Table 1 Clustering models

	1st-level clustering		2nd-level clustering	
Model	Algorithm	K_1 value	Algorithm	K_2 range
A	–	–	K-means	2–64
B	–	–	Dendrogram	2–64
C41	SOM	16	K-means	2–15
C42	SOM	16	Dendrogram	2–15
C51	SOM	25	K-means	2–24
C52	SOM	25	Dendrogram	2–24
C61	SOM	36	K-means	2–35
C62	SOM	36	Dendrogram	2–35
C71	SOM	49	K-means	2–48
C72	SOM	49	Dendrogram	2–48
C81	SOM	64	K-means	2–62
C82	SOM	64	Dendrogram	2–62

possible number of clusters. As 2nd-level clustering, both K-means and dendrogram were run with K_2 ranging between 2 and $K_1 - 1$ with unit pace.

To better understand the proposed procedure, the main points are summarized with reference to a fictional model Cad:

- As 1st-level clustering, a SOM is run with grid dimension set to a, so that the maximum allowed number of clusters is a^2. To minimize random errors, SOM is run 10 times and the best result is chosen as the one that minimizes the sum of distances data/cluster centroid. For each cluster, the centroid is computed as the mean of all the patterns in the cluster.
- As 2nd-level clustering, another clustering method (K-means if $d = 1$, dendrogram if $d = 2$) is run where the a^2 centroids are used as the input and K_2 is set to a value b ranging between 2 and $a^2 - 1$. K-means is run 10 times with $K_2 = b$ and for each run the algorithm replicates are set to 1000, in order to reach convergence and minimize the influence of initial points (same parameters were used for model A). Again, the best result is chosen as the one that minimizes the sum of distances data/cluster centroid. If the dendrogram is used, K_2 is set to b and there is no need to iterate computations, since the method is only based on initial distances.
- Finally, the original 168 patterns that were used as the input for 1st-level clustering are reassigned to the b new clusters, and DBI and CH can be computed with reference to the final partition, called “cluster solution”.

Figure 2 shows the variation of DBI and CH for each cluster solution of models in Table 1. Comparison between models A and B shows that model B is preferable since it has the lowest DBI and the highest CH for each cluster solution. Comparison of models A and B with models C shows that the use of a mixed strategy enhances clustering performance, resulting in lower DBI for each K_2 (this does not happen if CH values are observed). However, Fig. 2 shows that the estimate of both CVIs is somehow biased since, for each model, inspection of DBI and CH curves gives opposite results (the lowest DBI corresponds to low values of CH for most of the models). The reason for such a distortion lies in the observation that all the obtained cluster solutions are made up of a number of highly populated clusters, presumably related to significant consumers behaviors, along with sparsely populated clusters, probably related to behavior anomalies or measurement problems. Such a heterogeneity distorts the evaluation of the proposed indices because there is no way of emphasizing the higher significance of populated clusters.

In order to overcome such a bias and to successfully extract useful information from CVIs inspection, for each cluster solution of models C, only the clusters containing more than 5 patterns were considered and DBI and CH were recomputed. Figure 3 shows original and recomputed CVIs for models C71 and C72 as an example; considerations are similar for all the other models. It must be noted that for original CVIs K_2 coincides with the number of clusters in the cluster solution, whereas for recomputed CVIs K_2 must only be interpreted as a label for comparison purposes, and the actual number of clusters can be lower than or equal to K_2. As concerns DBI (Fig. 3a), recomputed values provide an optimal solution which is far more acceptable than the one provided by original DBI, since K_2 is now intermediate with respect to the extreme values. Moreover, the new optimal solution has a lower DBI value, because neglecting sparsely populated clusters emphasizes the meaningfulness of the remaining ones. Recomputed CH values (Fig. 3b) provide solutions

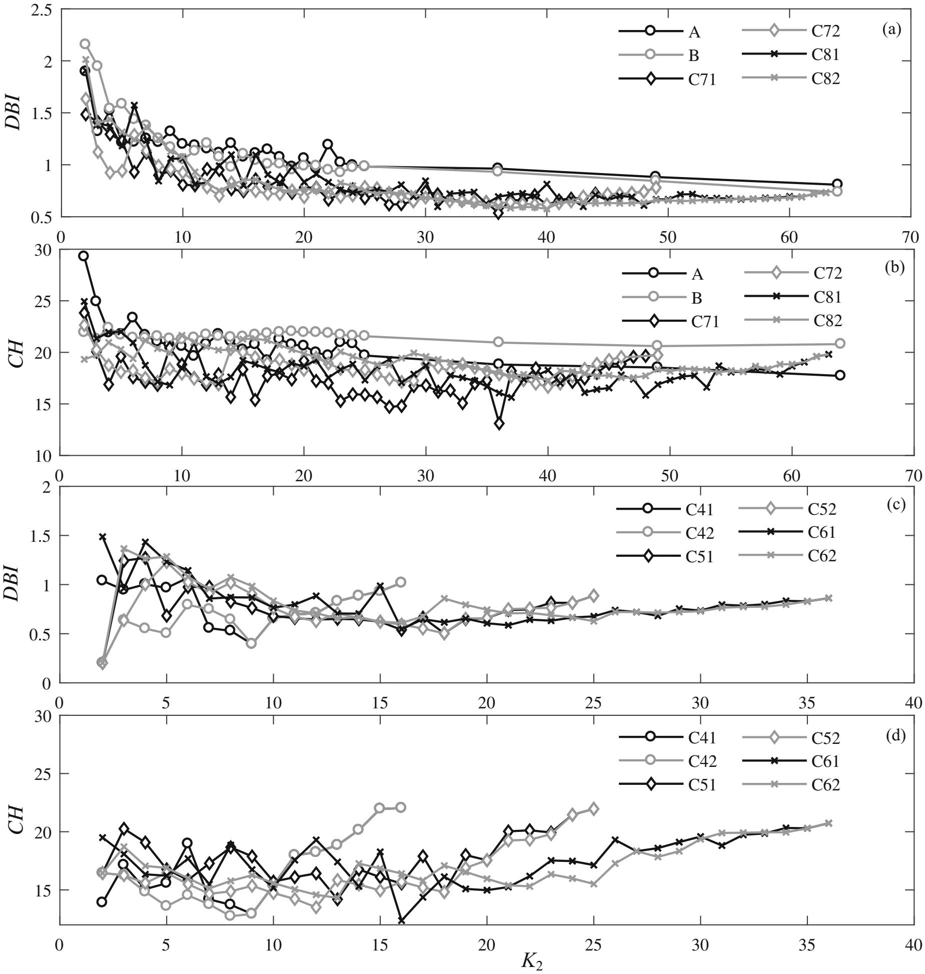


Fig. 2 Clustering evaluation for models in Table 1 in terms of **a**, **c** DBI and **b**, **d** CH

that are now coherent with those provided by the recomputed DBI; also, CH is now characterized by pronounced spikes highlighting clustering solutions that are more meaningful than the others.

A visual inspection of the best cluster solution for each model suggests that the optimal cluster solution is $K_2 = 31$ for model C71, which is characterized by 5 clusters containing more than 5 patterns. This solution is a relative maximum of recomputed CH (the absolute maximum values having an inconsistent number of meaningful clusters) and at the same time it corresponds to a recomputed DBI value which is very close to the minimum. Also, recomputed CH and DBI for this particular solution are among the maximum and minimum values, respectively, among all the solutions provided by the different tested models. Once the optimal solution was found, the 26 sparsely populated clusters were manipulated and suitably merged, reducing their number to 19 (5 highly populated plus 14 sparsely populated clusters).

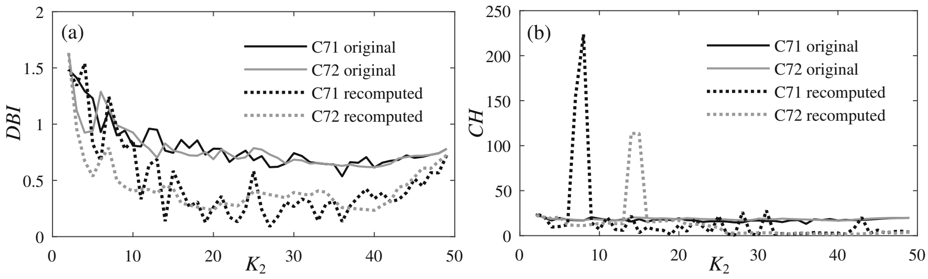


Fig. 3 Comparison between original and unbiased **a** DBI and **b** CH for models C71 and C72 in Table 1

Figure 4 shows the clusters of the optimal cluster solution as dark gray lines. Clusters 1–5 contain more than 5 pattern each, and they can be considered meaningful because they have correctly detected different customers behaviors; those clusters mainly differ for August consumptions, although minor differences exist in the remaining months. In cluster 2 standardized August discharge coincides with the maximum consumption of the year; this is usually justified with the observation that water demand can be correlated to air temperature (Alvisi et al. 2007), which is maximum in August at the latitude of interest. In cluster 1 August consumption is similar to the other months, resulting in a roughly horizontal trend. Moving from cluster 3 to cluster 5 August consumption progressively decreases, resulting in a downward spike which is more and more pronounced. This can be explained with the observation that people in Italy usually take their holidays in August, in a number of vacation days which can be related to their income level. Clusters 6–10 can be considered meaningful as well since they have detected anomalous behaviors probably caused by instrumental problems; alternatively, they can be considered representative of households which were not occupied for large periods of the year. Clusters 11–19 are singleton clusters; such patterns were not assigned to any of the other clusters in the clustering sample since their Euclidean distance from any of the centroids was larger than the average within distance in Eq. 6. Sparsely populated clusters will be discussed later on in the paper.

Clustering results were the inputs for supervised classification, applied to the validation sample made up of the remaining 2286 normalized monthly time series. It was found that including in the classification singleton or sparsely populated clusters caused a systematic decrease in the quality of training results. As a consequence, the neural network was trained using as input the annual patterns of the calibration sample belonging to clusters 1–5. Figure 5 shows the confusion matrix and the ROC curves resulting from the neural network training by means of a decision tree, which proved to be the most accurate model. Training results are highly satisfying: Fig. 5a shows that the TPR values are very high for all the five clusters, ranging from 100% for cluster 1 (which is also the most populated) to 68% of cluster 2 (related row of the confusion matrix shows that the model tends to overlap clusters 2 and 3). Figure 5b shows that, for each of the five clusters, the ROC curve is highly satisfying, having a AUC value ranging from 88% for cluster 3 to 100% for cluster 1; similarly, the CCP coordinates for each cluster are close to the optimal values (0,1).

The trained model was used to classify the remaining time series, that were assigned a cluster label from 1 to 5. Successively, a correction was manually operated to account for the neglected clusters, looking for those patterns that were classified 1–5 in the absence of any alternatives, when they actually belong to clusters 6–19. Specifically, for each pattern the Euclidean distance from the corresponding cluster centroid was computed; then, it was

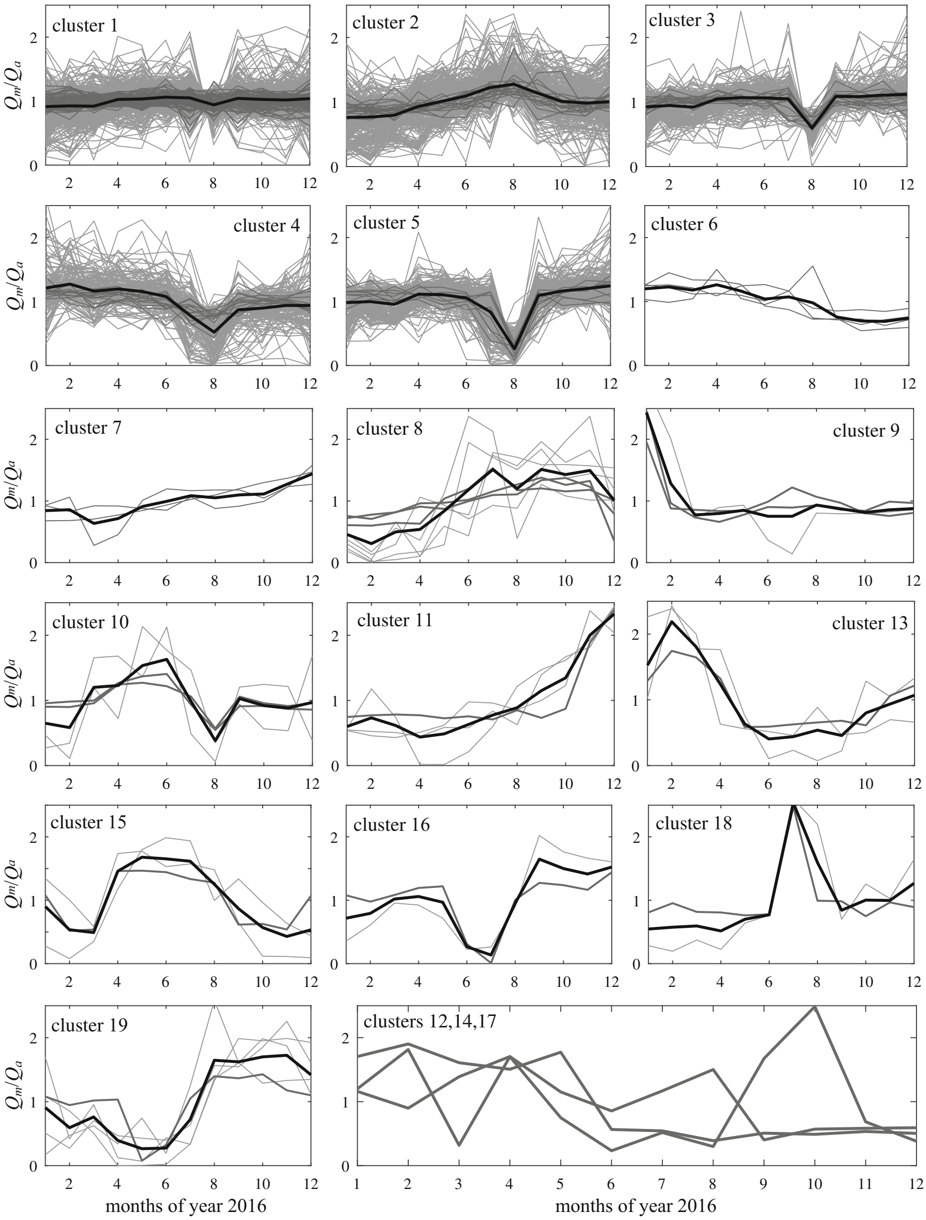
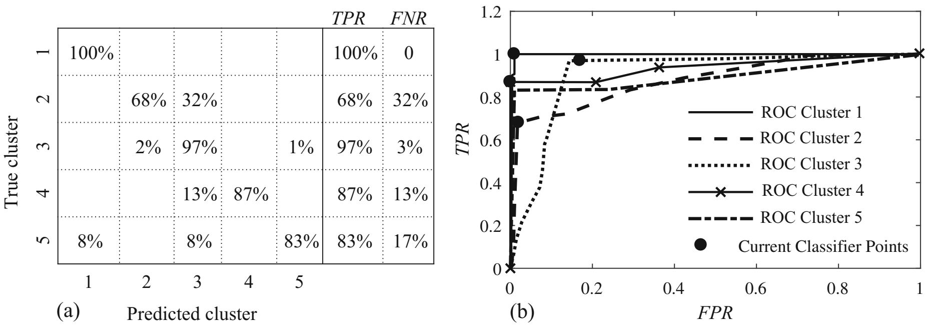


Fig. 4 Clusters provided by the best cluster solution for clustering (dark gray lines) and classification samples (light gray lines), and related centroids (black lines)

assumed that patterns having distances higher than 2 were wrongly labeled (a value of 2 was set as threshold since in each cluster in Fig. 4 the maximum Euclidean distance was lower than or equal to 1). For these patterns, Euclidean distances were computed with respect to the centroids of clusters 6–19 and pattern labels were changed if the smallest of these distances was lower than 2.



True cluster \ Predicted cluster	1	2	3	4	5	*TPR*	*FNR*
1	100%					100%	0
2		68%	32%			68%	32%
3		2%	97%		1%	97%	3%
4			13%	87%		87%	13%
5	8%		8%		83%	83%	17%

Fig. 5 Training results: confusion matrix (**a**) and ROC curves (**b**)

Figure 4 shows the final composition of clusters and clusters centroids. It is clear from Fig. 4 that classification confirms the consumption patterns detected by clustering, mainly highlighting the August downward plunge in consumption. Moreover, clusters 11, 13, 15, 16 and 18 (singleton in the clustering phase) are assigned additional patterns and can be discussed. Cluster 16 can be compared to cluster 5 in the general trend, although its minimum is reached in the month of July, which can be explained with a number of household occupants having anticipated their holidays. Cluster 10 can be compared to cluster 5 as well, although there is an abnormal consumption during spring. Clusters 6, 9 and 13 show a decreasing trend, with an elbow in October, March and June respectively, in opposition with the increasing trend shown by clusters 7 and 11. All the remaining clusters show an abnormal consumption in different months of the year.

The final repartition of patterns is the following: for the clustering sample, 54.17%, 11.36%, 8.93%, 7.14%, 4.76%, 2.38%, 1.79%, 1.78%, 1.19%, 1.19%, 0.59%, 0.59%, 0.59%, 0.59%, 0.59%, 0.59%, 0.59%, 0.59% and 0.59% patterns belong to clusters from 1 to 19, respectively; for the classification sample, 40.31%, 23.00%, 15.12%, 7.45%, 13.00%, 0, 0, 0.27%, 0.05%, 0.11%, 0.16%, 0, 0.11%, 0, 0.11%, 0.05%, 0, 0.05% and 0.21% patterns belong to clusters from 1 to 19, respectively. A general coherence can be observed between the two datasets, confirming that the clustering sample is suitably representative of the whole network.

4 Conclusions

In the present paper a procedure is presented to detect water consumption patterns describing significant consumers behaviors. The analysis is based on hourly consumption data collected within a large-size Smart Water Network located in Soccavo (Naples, Italy). Presented data are referred to single-household residential flow meters, connected to the Smart Water Network via a telemetry system, and collected in 2016. The procedure is structured in two consecutive phases ("clustering" and "classification") enabling the detection of 5 highly populated plus 14 sparsely populated clusters of annual patterns, the former group detecting significant consumption behaviors and the latter presumably related to instrumental errors or to households that were not occupied for prolonged periods. The methodology presents some novelty elements. The adoption of SOM, although frequent in the framework of electrical consumption, is very rare for water demand problems. Moreover, the case study represents a useful example of mixed clustering strategy; the final partitioning stems from an in-deep analysis of clustering parameters that, apparently, provide for contradictory information whose interpretation is not straightforward.

The research was strongly supported by the local water company, who provided for water consumption data, because of the high potentialities of results for both research and management purposes. Detected patterns will be of great aid in inferring about non-monitored connections; this will allow for: (i) a more realistic calibration of bills; (ii) the efficiency increase on the long term; (iii) the evaluation of water balance in the WDN and, as a consequence, the estimation of leakage volumes; (iv) the evaluation of the outputs at significant spatial scales, such as the census scale, with the aim of seeking correlations with socio-demographic parameters.

Acknowledgements The Authors would like to thank ABC Acqua Bene Comune – Napoli, who installed the telemetry system and provided for consumption data.

Compliance with Ethical Standards

Conflict of interests There is no conflict of interest.

References

Alvisi S, Franchini M, Marinelli A (2007) A short-term, pattern-based model for water-demand forecasting. J Hydroinf 9(1):39–50

Arbués F, Villanúa I, Barberán R (2010) Household size and residential water demand: an empirical approach. Aust J Agric Resour Econ 54(1):61–80

Avni N, Fishbain B, Shamir U (2015) Water consumption patterns as a basis for water demand modeling. Water Resour Res 51(10):8165–8181

Bergel T, Szelag B, Woyciechowska O (2017) Influence of a season on hourly and daily variations in water demand patterns in a rural water supply line–case study. J Water Land Dev 34(1):59–64

Blokker E, Vreeburg J, Van Dijk J (2010) Simulating residential water demand with a stochastic end-use model. J Water Resour Plan Manag 136(1):19–26

Bocci L, Mingo I (2012) Clustering large data set: An applied comparative study. In: Advanced statistical methods for the analysis of large data-sets. Springer, Berlin, pp 3-12

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. Chapman & Hall/CRC, Boca Raton

Briggs WM, Zaretzki R (2008) The skill plot: a graphical technique for evaluating continuous diagnostic tests. Biometrics 64(1):250–256

Browne AL, Medd W, Anderson B (2013) Developing novel approaches to tracking domestic water demand under uncertainty - A reflection on the “up-scaling” of social science approaches in the United Kingdom. Water Resour Manag 27(4):1013–1035

Caliński T, Harabasz J (1974) A dendrite method for cluster analysis. Commun Stat Theory Methods 3(1):1–27

Cominola A, Giuliani M, Castelletti A, Rosenberg DE, Abdallah A (2018) Implications of data sampling resolution on water use simulation, end-use disaggregation, and demand management. Environ Model Softw 102:199–212

Cousineau D, Chartier S (2010) Outliers detection and treatment: a review. Int J Psychol Res 3(1):58–67

Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other Kernel-based learning methods, 1st. Cambridge University Press, Cambridge

Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 2:224–227

Dimitriadou E, Dolničar S, Weingessel A (2002) An examination of indexes for determining the number of clusters in binary data sets. Psychometrika 67(3):137–159

Fawcett T (2004) ROC Graphs: notes and practical considerations for researchers. Mach Learn 31(1):1–38

Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27(8):861–874

Ferreira AM, Cavalcante CA, Fontes CH, Marambio JE (2013) A new method for pattern recognition in load profiles to support decision-making in the management of the electric sector. Int J Electr Power Energy Syst 53:824–831

Fontanazza CM, Notaro V, Puleo V, Freni G (2016) Multivariate statistical analysis for water demand modelling: implementation, performance analysis, and comparison with the PRP model. J Hydroinf 18(1): 4–22

Friedman JH, Bentley JL, Finkel RA (1977) An algorithm for finding best matches in logarithmic expected time. ACM Trans Math Softw 3(3):209–226

Gargano R, Tricarico C, Del Giudice G, Granata F (2016) A stochastic model for daily residential water demand. Water Sci Technol Water Supply 16(6):1753–1767

Ghavidelfar S, Shamseldin AY, Melville BW (2017) A multi-scale analysis of single-unit housing water demand through integration of water consumption, land use and demographic data. Water Resour Manag 31(7):2173–2186

Haque MM, de Souza A, Rahman A (2017) Water demand modelling using independent iomponent regression technique. Water Resour Manag 31(1):299–312

Johnson SC (1967) Hierarchical clustering schemes. Psychometrika 32(3):241–254

Jota PR, Silva VR, Jota FG (2011) Building load management using cluster and statistical analyses. Int J Electr Power Energy Syst 33(8):1498–1505

Kalteh AM, Hjorth P, Berndtsson R (2008) Review of the Self-Organizing Map (SOM) approach in water resources: analysis, modelling and application. Environ Model Softw 23(7):835–845

Keogh E, Chakrabarti K, Pazzani M, Mehrotra S (2001) Locally adaptive dimensionality reduction for indexing large time series databases. ACM SIGMOD Record 30(2):151–162

Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol Cybern 43(1):59–69

Krzanowski WJ (1988) Principles of multivariate analysis: a user's perspective. Oxford University Press, Clarendon

Laspidou C, Papageorgiou E, Kokkinos K, Sahu S, Gupta A, Tassiulas L (2015) Exploring patterns in water consumption by clustering. Procedia Engineering 119:1439–1446

Lebart L, Morineau A, Piron M (2004) Statistique exploratoire multidimensionnelle. Dunod, Paris

López JJ, Aguado JA, Martín F, Munoz F, Rodríguez A, Ruiz JE (2011) Hopfield–K-Means clustering algorithm: A proposal for the segmentation of electricity customers. Electr Power Syst Res 81(2):716–724

Loureiro D, Mamade A, Cabral M, Amado C, Covas D (2016) A comprehensive approach for spatial and temporal water demand profiling to improve management in network areas. Water Resour Manag 30(10):3443–3457

Macedo M, Galo J, De Almeida L, Lima AdC (2015) Demand side management using artificial neural networks in a smart grid environment. Renew Sust Energ Rev 41:128–133

MacQueen J et al (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA, vol 1, pp 281-297

McKenna S, Fusco F, Eck B (2014) Water demand pattern classification from smart meter data. Procedia Engineering 70:1121–1130

Papa JP, Falcao AX, Suzuki CT (2009) Supervised pattern classification based on optimum-path forest. Int J Imaging Syst Technol 19(2):120–131

Parker JM, Wilby RL (2013) Quantifying household water demand: a review of theory and practice in the UK. Water Resour Manag 27(4):981–1011

Popivanov I, Miller RJ (2002) Similarity search over time-series data using wavelets. In: Proceedings of the 18th international conference on data engineering, San Jose, CA, USA, pp 212-221

Powers DM (2007) Evaluation: from precision, recall and F-Factor to ROC, informedness, markedness & correlation. Tech. Rep. SIE-07-001, School of Informatics and Engineering Flinders University, Adelaide, Australia

Räsänen T, Voukantsis D, Niska H, Karatzas K, Kolehmainen M (2010) Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. Appl Energy 87(11):3538–3545

Sancho-Asensio A, Navarro J, Arrieta-Salinas I, Armendáriz-Íñigo JE, Jiménez-Ruano V, Zaballos A, Golobardes E (2014) Improving data partition schemes in smart grids via clustering data streams. Expert Syst Appl 41(13):5832–5842

Schikuta E (1996) Grid-clustering: an efficient hierarchical clustering method for very large data sets. In: Proceedings of the 13th international conference on pattern recognition, Wien, Austria, vol 2, pp 101–105

Verdú SV, García MO, Senabre C, Marín AG, Franco FG (2006) Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps. IEEE Trans Power Syst 21(4):1672–1682

Zhou Kl, Yang Sl, Shen C (2013) A review of electric load classification in smart grid environment. Renew Sust Energ Rev 24:103–110

Zhu X (2006) Semi-supervised learning literature survey. Computer Science Tech Rep 1530, University of Wisconsin-Madison