Water Resources Management (2019) 33: 739–755
https://doi.org/10.1007/s11269-018-2140-0

Pattern Detection and Scaling Laws of Daily Water Demand by SOM: an Application to the WDN of Naples, Italy

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

Received: 1 June 2018 / Accepted: 12 November 2018 /
Published online: 20 November 2018

Abstract
In the present paper, a novel method is provided to detect significant daily consumption patterns and to obtain scaling laws to predict consumption patterns for groups of homogeneous users. The first issue relies on the use of Self-Organizing Map to gain insights about the initial assumption of distinct homogeneous consumption groups and to find additional clusters based on calendar dates. Non-dimensional pattern detection is performed on both residential and non-residential connections, with data provided by one-year measurements of a large-size smart water network placed in Naples (Italy). The second issue relies on the use of the variance function to explain the dependence of aggregated variance on the mean and on the number of aggregated users. Equations and related parameters' values are provided to predict mean dimensional daily patterns and variation bands describing water consumption of a generic set of aggregated users.

Keywords Pattern detection · Scaling laws · Self-organizing map · Variance function · Water demand patterns

1 Introduction

Water demand modelling is a key issue in the context of an efficient management of Water Distribution Networks (WDNs) (Padulano and Del Giudice 2018). In order to obtain significant results in terms of water demand modelling, large consumption datasets are usually needed, that can be obtained by implementing the so-called "District Metering Areas" (DMAs), namely portions of a WDN connected to the rest of the network by a limited number of pipes (Gargano et al. 2016). In a DMA, all inflows and outflows are carefully monitored, with special focus on residential connections referring to single families,

✉ 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

which usually constitute the most part of the total number of connections in the DMA (Buchberger and Nadimpalli 2004). Consumption data can be collected at different time and space scales; accordingly, with increasing obtainable information they are usually classified in low-resolution, medium-resolution and high-resolution data (House-Peters and Chang 2011; Cardell-Oliver 2013; Cominola et al. 2015). However, the issue of aggregating/disaggregating water consumption data is highly debated, both concerning the temporal and the spatial scale (Magini et al. 2008; Cole and Stewart 2013; Vertommen et al. 2015).

Water demand modelling has been performed with different approaches (House-Peters and Chang 2011; Padulano and Del Giudice 2018). Techniques and methods involve, among others, probabilistic analysis of time series (Zhou et al. 2002; Alvisi et al. 2007; Tricarico et al. 2007; Wong et al. 2010; Quevedo et al. 2010; Adamowski et al. 2012; Chen and Boccelli 2014; Hutton and Kapelan 2015; Gargano et al. 2016) and multivariate analysis (Mamade et al. 2014; Fontanazza et al. 2016; Wa'el et al. 2016; Loureiro et al. 2016; Cheifetz et al. 2017; Ghavidelfar et al. 2017; Haque et al. 2017), in the attempt to recognize seasonal cycles and to correlate water demand to significant weather (temperature, rainfall, evapotranspiration) and socio-demographic (building and billing information, age and education of consumers) variables. Although classical statistical methods are still largely used, adoption of innovative techniques such as Kalman filters (Nasseri et al. 2011), Artificial Neural Networks (Firat et al. 2010; Adamowski et al. 2012; Bennett et al. 2013) and Self-Organizing Maps (Padulano and Del Giudice 2018) is becoming more and more frequent.

In the present paper, a procedure is proposed to detect daily water demand patterns within different classes of users and to use these findings to gain insights about spatial aggregation problems. For the first issue, Self-Organizing Map is used to obtain information about the structure of daily patterns for both residential and non-residential users; for the second issue, the theoretical framework by Magini et al. (2008) and Vertommen et al. (2015) is adopted and suitable scaling laws are calibrated. The procedure is validated with data collected within the DMA of Soccavo (Naples, Italy); part of this database was analysed by Padulano and Del Giudice (2018), although with different purposes.

2 Materials and Methods

2.1 Spatial Aggregation

Spatial aggregation of water consumption is a highly debated issue in the research community (Vertommen et al. 2015; Alvisi et al. 2015). The main concerns focus on the changes in the statistical moments of consumption time series that occur when a number of consumers are aggregated. In turn, these modifications alter the demand uncertainty, which should be carefully modelled when, for example, consumption profiles are assumed at the nodes of a water network model, reproducing the consumption of a given population for hydraulic simulation purposes (Buchberger and Wu 1995). Relations explaining the aggregated moments as a function of the number of aggregated users are often referred to as "scaling laws" (Vertommen et al. 2015).

Kottegoda and Rosso (2008) give the mathematical definition of the mean and the variance of an aggregated random variable, namely a random variable which is the sum of a number N of random variables. Vertommen et al. (2015) and Alvisi et al. (2015) suggest that, for water demand modelling purposes, hourly recorded data of water consumption registered at each different hour of the day constitute a separate random variable and should

be treated accordingly. Let $q_{h,i}(d)$ be defined as the random variable which describes the water volume consumed by a single household i within hour h of day d; if the observation time consists in a number of days g, the recorded sample for hour h will be made up of a maximum of g data, given that in practical situations some data could be missing due to, for example, transmission problems. For a group of N households, the aggregated volume at hour h of day d is given by:

$$Q_h(d) = \sum_{i=1}^{N} q_{h,i}(d) \qquad h = 1, \ldots 24 \tag{1}$$

which is a time series referring to a specific hour h, made up of one data point for each day d (so that the total number of data in the time series is g). The mean of the aggregated random variable $Q_h(d)$ (namely the mean of the time series across the available days g), henceforth called "aggregated mean" μ_h, is the sum of the means $\mu_{h,i}$ of the variables referring to each single household:

$$\mu_h = E\{Q_h\} = E\left\{\sum_{i=1}^{N} q_{h,i}\right\} = \sum_{i=1}^{N} E\{q_{h,i}\} = \sum_{i=1}^{N} \mu_{h,i} \tag{2}$$

with E being the expected value of the time series $Q_h(d)$, one for each hour h of the day, each time series made up of g values. If only single-family residential households are considered, water demand can be assumed homogeneous in space; this assumption implies that the users belonging to this group have similar behaviour in terms of water usage (Magini et al. 2017). In this case, it can be demonstrated that the mean μ_h is strictly related to the mean value $\mu_{h,u}$ computed across all the available households (Magini et al. 2017):

$$\mu_h = \sum_{i=1}^{N} \mu_{h,i} = \frac{1}{N} \cdot N \cdot \sum_{i=1}^{N} \mu_{h,i} = N \cdot \left(\frac{1}{N} \sum_{i=1}^{N} \mu_{h,i}\right) = N \cdot \mu_{h,u} \tag{3}$$

where $\mu_{h,u}$ will be henceforth called "single-user mean", since it constitutes a representative value for the generic household. Equation 3 highlights the linear dependence between the aggregated mean μ_h and the number of aggregated households N. Similarly, the variance of the aggregated variable $Q_h(d)$, henceforth called "aggregated variance" σ_h^2, is the sum of single household variances $\sigma_{h,i}^2$ corrected by a term accounting for the covariance among the households:

$$\begin{aligned} \sigma_h^2 = Var\{Q_h\} = Var\left\{\sum_{i=1}^{N} q_{h,i}\right\} &= \sum_{i=1}^{N} Var\{q_{h,i}\} + 2\sum_{i=1}^{N}\sum_{j>i}^{N} Cov\{q_{h,i}, q_{h,j}\} \\ &= \sum_{i=1}^{N} \sigma_{h,i}^2 + 2 \sum_{i,j=1, i\neq j}^{N} Cov_{h,ij} \end{aligned} \tag{4}$$

Similarly to the aggregated mean, if water demand is assumed a homogeneous process, aggregated variance can be found to be strictly related to the "single-user variance" $\sigma_{h,u}^2$, computed as the mean value of the variances across all the available households (Vertommen et al. 2015). Introduction of single-user variance in Eq. 4 provides:

$$\sigma_h^2 = N \cdot \sigma_{h,u}^2 + 2 \sum_{i,j=1, i\neq j}^{N} Cov_{h,ij} \tag{5}$$

where the dependence of the aggregated variance on the number of aggregated households N is not explicated. Only in the absence of any spatial correlation among the household demands, namely if the covariance term is null, the aggregated variance can be found to be linearly dependent on N (Vertommen et al. 2015). In any other case, the dependence can be put in the generic form

$$\sigma_h^2 = \alpha \cdot N^\beta \tag{6}$$

where the coefficients α and β depend on the structure of the spatial correlation and on $\sigma_{h,u}^2$ (Magini et al. 2008). Considering the linear dependence between μ_h and N provided by Eq. 3, Eq. 6 can be rewritten as follows:

$$\sigma_h^2 = \gamma \cdot \mu_h^\beta \tag{7}$$

with $\gamma = \alpha / \mu_{h,u}^\beta$. The dependence between the mean and the variance of a sample, such as that expressed by Eq. 7, is usually referred to as "variance function" (Davidian and Carroll 1987). Variance function plays a key role in the framework of Generalized Linear Models. Indeed, the linear model in Eq. 3 can be regarded to as the systematic component of a Generalized Linear Model with identity link function (McCullagh and Nelder 1989), relating the aggregated mean of consumption (the response variable) with the number of aggregated users (the explanatory variable). Equation 7 is the variance function related to the model, whose structure suggests that the regression model is heteroscedastic, since only if $\beta = 0$ the variance would be independent on the mean. This is coherent with the consideration that the assumptions of normality and homoscedasticity are often violated in nature (Madsen and Thyregod 2010), for example for physically bounded data resulting in highly skewed distributions, which could be the case of water consumption (Buchberger and Nadimpalli 2004).

2.2 Experimental Water Demand Database

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 4253 customer connections whose flow meters were completely replaced during the last three years. Figure 1 and Table 1 show the repartition of flow meters among residential, commercial, institutional or service-related flow meters (hereinafter shortly called "commercial").

Data were recorded in the period January 1^{st} to December 31^{st}, 2016. Each data represents the overall water volume consumed by all the unknown appliances located downstream of the measuring flow meter within the hour preceding the measure, in litres; the consumed volume is assumed constant within the hour, so that recorded water consumption can be expressed as a discharge, in litres per hour. Data underwent a preliminary cleansing procedure to detect outliers and discard unreliable time series; this caused the elimination of a number of time series (Padulano and Del Giudice 2018) and prevented some types of connections (i.e. food stores and medical centres among others) from further analyses. The final number of analysed time series, along with the labels for the consumption groups, can be find in Table 1 (2002 residential and 104 commercial connections). For residential

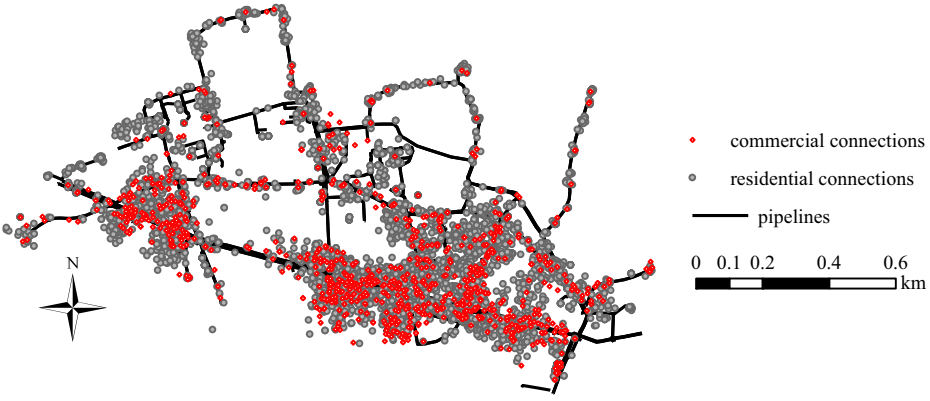

Fig. 1 District metering area in Soccavo (Naples, Italy)

Table 1 Water consumption database features

connection description	N (DMA)	N^* (analyzed)	group	$E\{\mu_d\}$ (L/h)	$E\{\mu_{d,u}\}$ (L/h)
single households (cluster 1)	2998	848	1	15352	18.5
single households (cluster 2)		451	2	6742	15.2
single households (cluster 3)		299	3	4862	16.6
single households (cluster 4)		252	4	3962	16
single households (cluster 5)		152	5	2445	16.2
single households (other clusters)		44	–	–	–
multiple households	182	–	–	–	–
unknown residential flow meters	485	–	–	–	–
clothing shops	198	47	6	303	6.34
garages	62	–	–	–	–
food stores	48	–	–	–	–
medical centers	29	–	–	–	–
barber shops	28	21	7	371	39.6
coffee shops	19	16	8	635	17.9
schools	5	3	9	310	96.8
banks and postal offices	7	4	10	74	19.1
offices	47	13	11	213	17.2
laundries	11	–	–	–	–
beauty centers	11	–	–	–	–
restaurants	9	–	–	–	–
supermarkets	6	–	–	–	–
sporting centers	6	–	–	–	–
rail stations	4	–	–	–	–
others	98	–	–	–	–

*data belonging to the group of interest and passing data reliability analysis

single-household flow meters an additional repartition is shown in Table 1 due to a significantly different behaviour in the water volumes consumed in the month of August, as found in Padulano and Del Giudice (2018).

Table 1 shows, for each group, the aggregated daily discharge Q_d averaged across the available days of measurements within 2016, as a reference value (here called $E\{Q_d\}$). As expected, for the residential groups, $E\{Q_d\}$ is highly dependent on the number of aggregated users, descending from group 1 to group 5. For the commercial groups, groups 8 and 7 show the highest average daily consumption, consistently with the significant water usage typical of those activities, whereas the remaining groups show a water usage which decreases with the number of possible restroom/canteen users (a reasonable assumption is pupils, students and teachers for group 9, employees and customers for group 11, employees for group 10). Table 1 also shows the single-user daily discharge $\mu_{d,u}$ averaged across the available days of measurements within 2016, as a reference value (here called $E\left\{\mu_{d,u}\right\}$); it is interesting to note that, for residential users, differently from $E\{Q_d\}$, $E\left\{\mu_{d,u}\right\}$ shows no significant dependence on the number of users in each cluster; this can be explained by the consideration that each residential cluster gathers households with different number of occupants, having similar average consumption pattern but different absolute values. As regards non-residential users, schools show the highest average daily consumption, followed by barber shops.

3 Discussion of Results

3.1 Detection of Consumption Prototypes

The hypothesis of homogeneous demand implies that several "consumption groups" exist; each consumption group gathers consumers with similar behaviour in terms of water consumption, so that a representative pattern, henceforth called "prototype", at significant time scales can be predicted. For instance, a typical daily pattern describing water consumption of a residential user is expected to be different from the pattern related to a commercial activity in terms of peak time, peak entity and base value. In turn, it is possible that two residential consumption patterns differ if they refer to families with significantly different number of components or usage habits of water appliances.

In order to check the existence of different water demand prototypes, the Self-Organizing Map was adopted (Kohonen 1982; Verdú et al. 2006; Kalteh et al. 2008; Padulano and Del Giudice 2018). This choice was made to let the profiling be completely unsupervised (since SOM needs no preliminary information), with the only initial assumption about consumption groups (Table 1). It must be noted that, even if a set of homogeneous patterns is considered (for example belonging to single households with the same number of occupants), those patterns will be characterized by deep fluctuations both in time (if many days of hourly data are compared for the same user) and in space (if consumption data at the same hour are compared for multiple users). This is due to erratic, unpredictable events such as a temporary vacation in the household or a day off from work. To enhance the performance of pattern recognition a typical approach is to aggregate water demand within each consumption group; in other words, the input of SOM consists in a time series which is the sum of the time series belonging to each group. This was considered acceptable since information about the number of components for each family was not systematically available in the WDN.

3.1.1 Residential Consumption

Table 1 shows the consumption groups used as input for SOM. For each group, the aggregated time series, representing the total water consumption of the group, was timely aggregated at the daily scale. Then, for each group the time series at the hourly scale was cleansed of any seasonal effect by dividing each hourly discharge Q by the corresponding daily value Q_d and data were organized in $g \times 24$ matrices (for each group g is the number of available days of measurement, ≤ 366). For each group the corresponding matrix was used as input for SOM, whose grid dimension was set to 20, so that the final output neuron grid is 20×20. A very large grid dimension was set because different labels for data were available, as will be discussed in the following paragraphs. Figure 2 shows the results of SOM for the 5 residential consumption groups, with different labelling strategies.

The notation 1–7 with 1 = Monday and 7 = Sunday (Fig. 2a, c, d, e, f) was used to test the possible differences in the non-dimensional daily pattern between workdays and holidays (weekly cycle). For residential consumption groups, 3 clusters can be obtained: Cluster A (Sundays), Cluster B (Saturdays) and Cluster C (Monday–Friday). Clusters A and B also contain 13 public holidays which SOM automatically chose to place in Cluster A or in Cluster B. Moreover, some other days were labelled as "optional holidays" because for some users they may be normal workdays; these are the days between Christmas and Epiphany and the days immediately preceding and following Easter, when schools are closed and working people often choose to take some extra days off.

Calendar date labels (Fig. 2b) were used to test the possible differences in the non-dimensional daily pattern across the year (seasonal cycle); such labels identify the season (1 = spring,...4 = winter) each day belongs to. Figure 2b shows that there is no strong clustering in terms of climate, since days belonging to the same season are randomly distributed within the SOM (only a feeble tendency of summer days to agglomerate can be observed). This implies that, although time series were initially divided into 5 different clusters basically according to the total consumption in August (Padulano and Del Giudice 2018), this difference can only be observed in dimensional patterns but it does not extend to the daily distribution of non-dimensional consumption.

Since SOM preserves feature topology, the position of neurons enables the prediction of variability within each cluster. For instance, for all the residential groups, variance in Cluster C is expected to be low, since there is a large number of days very close in the map, with few voids. On the opposite, Cluster A is expected to have a higher variance since it is made up of a reduced number of days scattered in a wide portion of the map, with a great number of empty neurons in it. Some other considerations can be made by merely observing the scattering of days in the map: in Fig. 2e, as an example, positions in Cluster A suggest that there is a large number of similar Sundays, placed in the upper part of the cluster, whereas a small number of isolated Sundays, placed in the lower left part of the map, is significantly different and separated from the first group by many empty neurons. Similarly, in Fig. 2c two distinct groups of Sundays can be detected. Cluster C in Fig. 2e is twofold as well, being made up of a main upper part of close neurons plus a small number of isolated days in the lower part (the empty neurons in-between suggesting different-looking patterns). The shape of clusters can give information about variability as well: for example, the shape of Cluster B is compact in Fig. 2d, e, f, suggesting a very low inner variability (since many days are connected to the others by all the sides of the hexagons) but it is extended in Fig. 2a, c, suggesting a slightly higher variability, since most days are only connected to the others by means of the upper and lower sides of the hexagon.

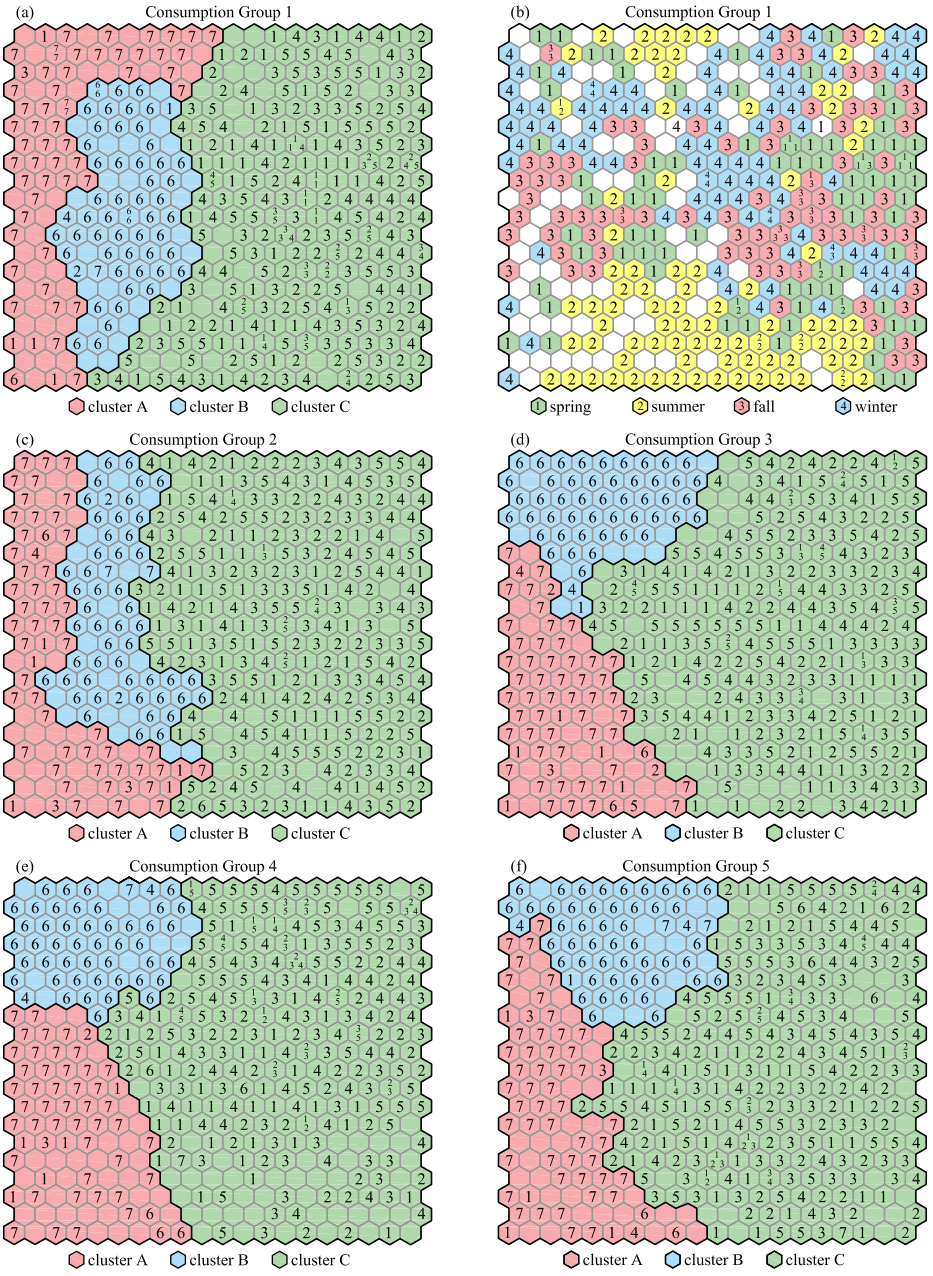

Fig. 2 SOM for consumption groups 1–5 (residential flow meters)

Figure 3 shows the final non-dimensional prototypes for residential consumption. The three clusters share some characteristics, such as the highest peak in the morning and a very low consumption in the night, with only a small variance. Moreover, Clusters A and B have a similar peak in the evening, which is significantly lower than the evening peak in Cluster

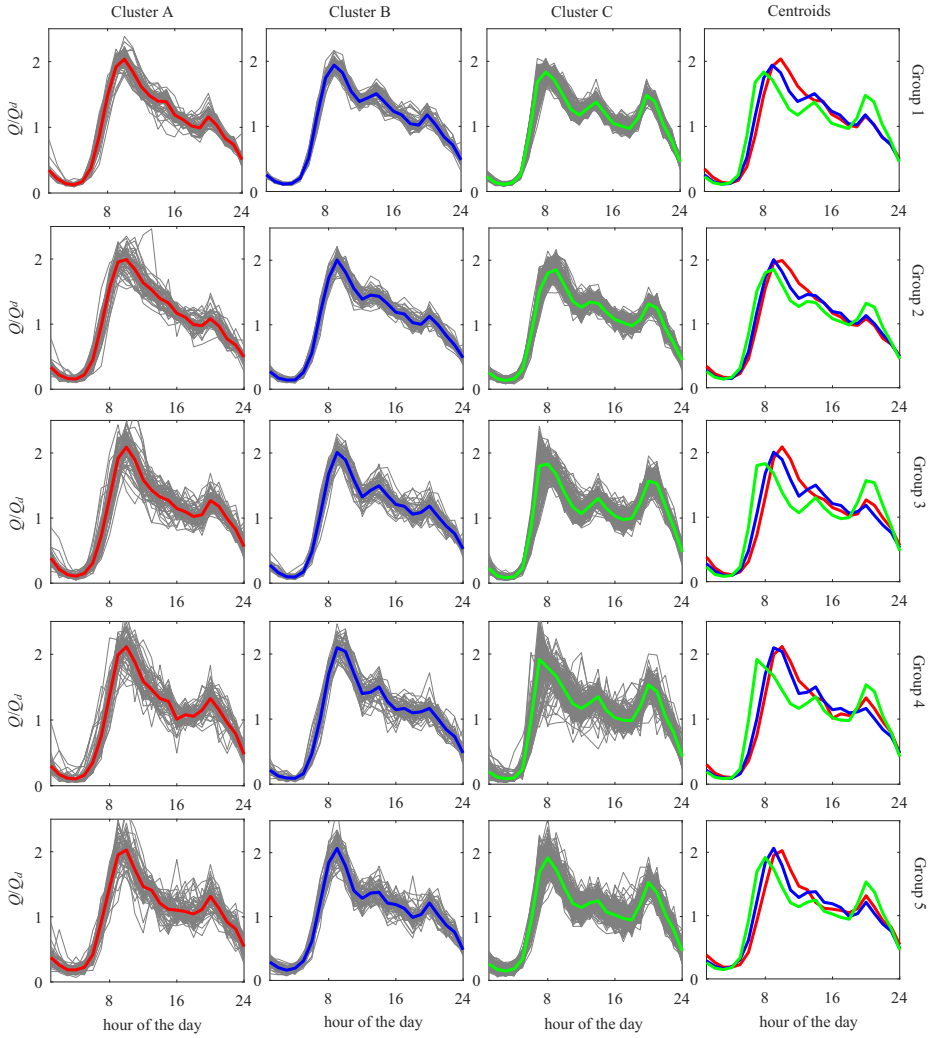

Fig. 3 Daily non-dimensional prototypes for consumption groups 1–5 (clustered daily patterns in first to third column, cluster centroids in fourth column)

C. In Clusters A and B the morning peak is equally delayed compared to Cluster C, whereas in Clusters B and C a third peak at lunchtime can be observed that is missing in Cluster A.

3.1.2 Commercial Consumption

Figure 4a shows the results of SOM for the commercial consumption group labelled "shops", which includes clothing and gift shops. SOM provides 3 clusters, namely: Cluster A, made up of highly scattered Sundays suggesting a significant variability; Cluster B, made up of grouped Saturdays; Cluster C, made up of workdays.

Figure 4b shows the results of SOM for the commercial consumption group of barber shops. SOM provides 3 clusters, organized in a peculiar shape. Cluster E, made up of days

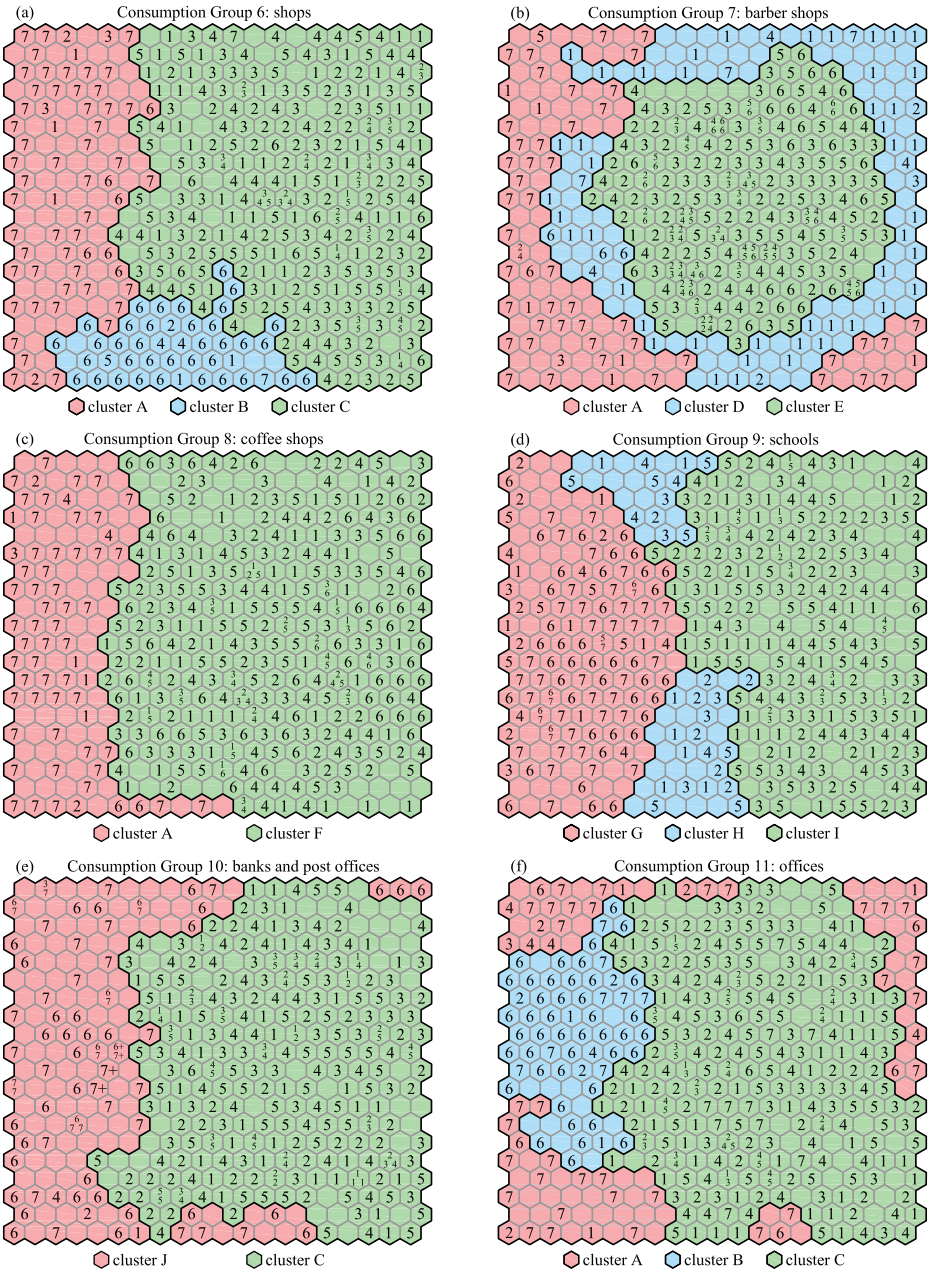

Fig. 4 SOM for consumption groups 6–11 (commercial flow meters)

from Tuesday to Saturday, is compressed in the middle of the map, suggesting both a high similarity of the patterns within the cluster and the need for SOM to arrange additional space to account for the variability of the other clusters. Cluster D is made up of isolated

Mondays: this is because for this day barber shops can autonomously decide whether to stay closed or half-day open. Cluster A is made up of dispersed Sundays.

Figure 4c shows the results of SOM for the commercial consumption group of coffee shops. Resulting clusters are Cluster A (Sundays) and Cluster F (days from Monday to Saturday). Along with barber shops, this is the only case when Saturdays are randomly scattered among the workdays.

Figure 4d shows the results of SOM for schools. For this group resulting clusters gather different sets of days, namely: Cluster G is made up of days when schools are completely closed (Saturdays, Sundays, public holidays and the greatest part of August, whose days collect in the first column of the map); Cluster H is made up of days when there are no lessons but some other activities could take place, such as meetings and exams (part of June and September, the whole July and Christmas holidays); Cluster I is made up of regular school days.

Figure 4e shows the resulting two clusters provided by SOM for banks and postal offices: Cluster C gathers regular workdays covering the whole year, whereas Cluster J gathers all Saturdays, Sundays and public holidays. In this cluster some neurons exist which are occupied by a great number of days, labelled as "7 +". This can be explained with the observation that, differently from other offices or stores, there is no possibility at all that banks and postal offices are in operation on Sundays.

Figure 4f shows the results of SOM for the connections labelled as "offices". The three resulting clusters are quite similar to those in Fig. 4a, separately gathering Sundays, Saturdays and workdays respectively, with public holidays assigned to Cluster A or Cluster B. For this consumption group, however, a larger variability among Sundays can be observed, represented by a partition of Cluster A in the four corners of the map.

Figure 5 shows the final non-dimensional prototypes for commercial consumption. Compared to Fig. 3, a significantly larger variability can be observed for all the consumption groups and especially for clusters excluding regular workdays. This is mainly due to (i) the small number of connections in each group causing a residual instability in consumption values and (ii) the occurrence of isolated water usages during the day, causing very high peaks in the non-dimensional pattern. It can also be noted that only consumption groups 9 and 10, whose opening is strictly allowed only during workdays, show zero consumption, with the non-dimensional pattern fluctuating around 1.

3.2 Scaling Laws for Residential Consumption Groups

In the present section, spatial aggregation issues are examined in depth for residential consumption groups only, given the high number of connections. Section 3.1 showed that no differences can be observed in the behaviour of non-dimensional patterns of groups 1–5; since those clusters mainly differ for consumption volumes in the month of August, it was decided to treat them as a unique consumption group (structured in Clusters A, B and C), provided that August observations are removed from the residential time series. The total number of time series belonging to groups 1–5 is $N_{\max} = 2002$.

With the aim of deriving dimensional patterns (mean and variance) for different spatial aggregation levels N (with $N = 5, \ldots 2000$ with pace 25, plus $N_{\max}$), the following procedure is adopted:

1. $\forall N$, N residential flow meters are randomly selected from the dataset and related time series $q_i(t)$ are summed up to obtain aggregated signals $Q(N, t)$. Random extraction is repeated 50 times and averaged to overcome the influence of specific connections.

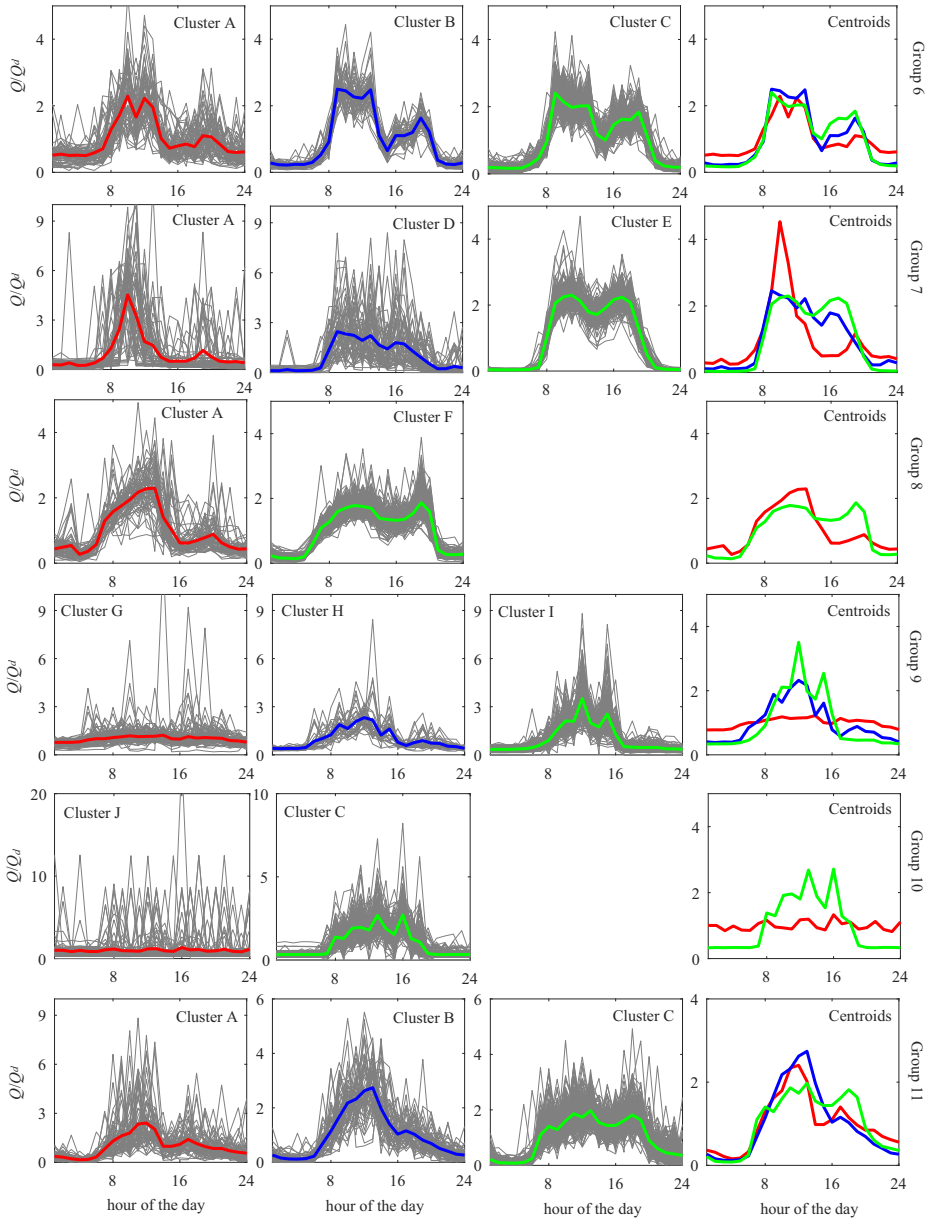

Fig. 5 Daily non-dimensional prototypes for consumption groups 6–11 (clustered daily patterns in first to third column, cluster centroids in fourth column)

2. $\forall N$, from the aggregated time series $Q(N,t)$ observations are extracted referring to a specific hour h of the day, to obtain hourly subseries $Q_h(N)$. Corresponding aggregated mean $\mu_h(N)$ and variance $\sigma_h^2(N)$ are computed and coefficients γ and β are calibrated by fitting Eq. 7 to data $Q_h(N)$.

3. The operations described at steps 1 and 2 are repeated, separately, for each cluster k (with k = A, B or C) and for each hour h of the day; therefore, the procedure is applied $3 \times 24 = 72$ times.

Figure 6a shows that the relation between μ_h and N is perfectly linear for each hourly subseries, in accordance with Eq. 3, for the three different clusters of days. Table 2 shows the angular coefficient $\mu_{h,u}$, which coincides with the mean of $q_{h,i}$, averaged across all the residential flow meters. If the $\mu_{h,u}$ column is ideally ordered from the highest to the lowest value, the order of the lines in Fig. 6a is also obtained, since for a fixed N the highest is $\mu_{h,u}$ the highest is the line: this implies that, for Cluster A, the highest line in Fig. 6a refers to $h = 10$, the second highest to $h = 9$ and so on to the lowest, which refers to $h = 4$; for Cluster B the highest line refers to $h = 9$, the second highest to $h = 10$ and the lowest to $h = 3$; for Cluster C the highest line refers to $h = 8$, the second highest to $h = 7$ and the lowest to $h = 3$ (Table 2). Figure 6b shows the relation between σ_h^2 and μ_h in accordance with the variance function in Eq. 7. Table 2 shows the coefficients γ and β for each cluster k and each hour h, obtained by a fitting operation on data. It is particularly evident that, although a slight variation exists, β can be considered constantly equal to 2, implying a perfect cross-correlation among the time series and confirming the hypothesis of homogeneous

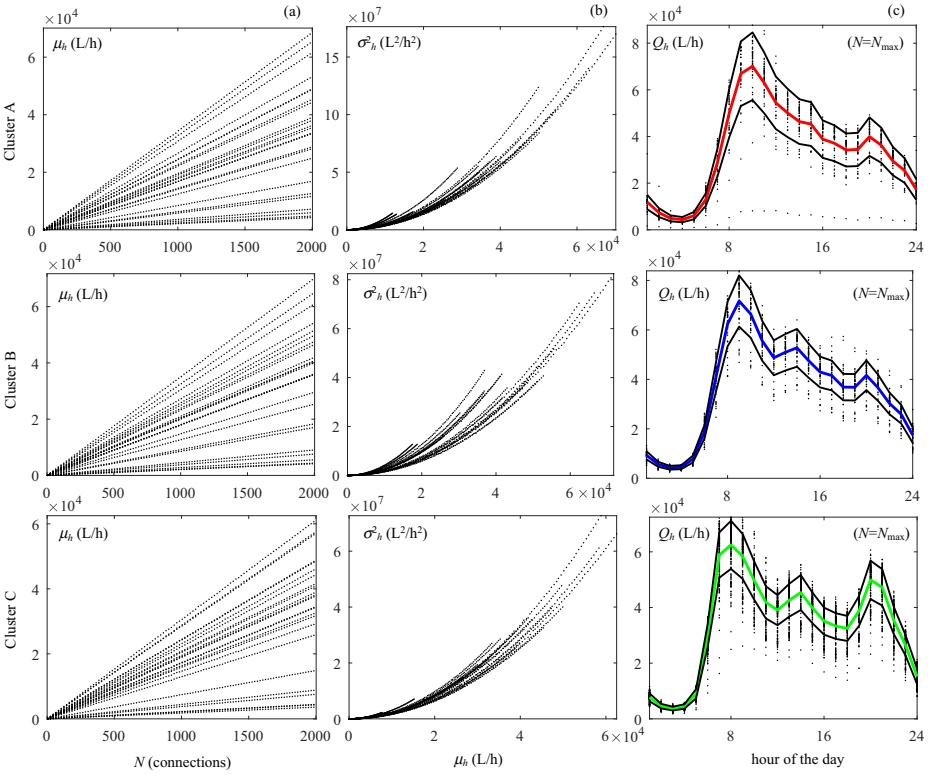

Fig. 6 Relation between **a** aggregation level N and aggregated mean μ_h data (fitted line parameters in Table 2, not shown) and **b** aggregated mean μ_h and variance σ_h^2 data (fitted curve parameters in Table 2, not shown); **c** aggregated data Q_h for each hour of the day at the maximum aggregation level $N_{\max} = 2002$, along with aggregated mean μ_h and $\pm\sigma_h$ variation band

Table 2 Regression parameters

	Cluster A			Cluster B			Cluster C		
h	$\mu_{h,u}$*	γ*	β*	$\mu_{h,u}$*	γ*	β*	$\mu_{h,u}$*	γ*	β*
1	5.8	0.104	2.00	4.5	0.035	2.00	3.8	0.040	1.99
2	3.6	0.093	2.00	2.8	0.032	1.99	2.2	0.041	1.99
3	2.4	0.052	2.00	2.1	0.036	1.98	1.8	0.038	1.99
4	2.1	0.053	1.99	2.2	0.023	1.99	2.2	0.026	1.99
5	3.0	0.056	1.99	3.8	0.031	1.99	4.4	0.027	2.01
6	6.3	0.082	2.00	9.2	0.038	2.00	14.6	0.021	2.00
7	14.1	0.067	2.00	20.0	0.025	2.00	28.8	0.021	2.00
8	24.4	0.051	2.00	30.5	0.018	2.00	30.5	0.017	2.00
9	32.7	0.040	2.00	35.0	0.017	2.00	28.5	0.018	2.00
10	34.3	0.035	2.00	32.5	0.017	2.00	24.3	0.019	2.00
11	30.8	0.036	2.00	27.2	0.018	2.00	20.5	0.021	2.00
12	26.6	0.035	2.00	23.8	0.017	2.00	19.1	0.021	2.00
13	24.5	0.034	2.00	24.9	0.015	2.00	20.8	0.019	2.00
14	22.7	0.035	2.00	25.8	0.015	2.00	22.2	0.016	2.00
15	22.2	0.039	2.00	23.3	0.017	2.00	19.5	0.017	2.00
16	19.0	0.042	2.00	21.0	0.020	2.00	17.2	0.021	2.00
17	18.1	0.044	2.00	20.3	0.024	2.00	16.3	0.025	2.00
18	16.7	0.047	2.00	18.0	0.033	2.00	15.9	0.021	2.00
19	16.8	0.043	2.00	18.0	0.026	2.00	18.9	0.017	2.00
20	19.5	0.038	2.00	20.3	0.018	2.00	24.4	0.016	2.00
21	17.7	0.039	2.00	17.9	0.026	2.00	23.1	0.018	2.00
22	14.4	0.040	2.00	14.8	0.029	2.00	17.3	0.023	2.00
23	12.4	0.041	2.00	12.7	0.026	1.99	12.9	0.022	2.00
24	8.4	0.050	2.00	8.5	0.043	2.00	7.4	0.032	2.00

*Variables units: $\mu_{h,u}$ (L/h); β (-); γ $(\text{L/h})^{2-\beta}$ (if $\beta = 2$ γ is non-dimensional)

demand within the group; in this sense, the assumption of a linear(quadratic) law to describe aggregated mean(variance) is perfectly validated by the experimental data, providing fitting regression coefficients higher than 98% for all the hours of the day. Similarly, variability of γ can be reduced to two different values γ_1 and γ_2 representative of nocturnal and diurnal hours respectively, with $\gamma_1 = E\{\gamma(h = 1 - 5, 24)\}$ and $\gamma_2 = E\{\gamma(h = 6 - 23)\}$. Table 2 provides $\gamma_1 = 0.068, 0.033, 0.034$ and $\gamma_2 = 0.044, 0.022, 0.020$ for k =A, B, C respectively. In Fig. 6b, the correspondence between curves and hours can be found by noticing that the shortest curves refer to the lowest consumption hours, whereas the longest refer to the peak hours. Figure 6c shows the final dimensional aggregated data for the maximum possible aggregation level ($N_{\max} = 2002$) along with aggregated mean μ_h and variation band $\pm\sigma_h$ (here σ_h is computed by means of γ_1 and γ_2). Two observations can be made:

1. Standard deviation for nocturnal hours is significantly lower than diurnal hours. This was expected since night consumption usually is very low and with little fluctuation in water usage behaviour of consumers;

2. For each hour of the day, the position of the mean in Fig. 6c with respect to aggregated data suggests a certain asymmetry in the probability density function of the hourly series (this is particularly evident for Cluster C).

4 Conclusions

In the present paper a novel approach is presented to detect daily water consumption patterns for both residential and commercial users, based on the use of Self-Organizing Map. As a novelty element in the framework of water demand, SOM was used to perform pattern recognition and to understand consumption prototypes. SOM results confirmed the existence of different "consumption groups", referred to different types of consumers, characterized by significantly different non-dimensional water consumption patterns; for each group, patterns can be further classified in clusters based on the weekdays they refer to (workdays/weekends with some differences among the groups). However, in each cluster, calendar date is irrelevant, meaning that non-dimensional daily patterns do not show any seasonal effect.

Basing on the homogeneous groups resulting from SOM and on real data, an in-deep analysis of residential prototypes was performed to solve the issue of spatial aggregation. As a novelty element, variance function was used to describe the dependence of aggregated variance on the number of aggregated users. Equations and parameters' values were given to generate mean dimensional consumption pattern and variation bands describing water consumption of a group of any N users with N ranging between 5 and 2002. The so-obtained patterns can be used for different purposes, such as: (i) as nodal inputs of a water distribution network to run hydraulic simulations; (ii) as design input for a supply line; (iii) to represent water consumption of significant sets of people, such as agglomerates of buildings and census areas.

Acknowledgments 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

Adamowski J, Fung Chan H, Prasher SO, Ozga-Zielinski B, Sliusarieva A (2012) Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. Water Resour Res 48(1):W01528

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

Alvisi S, Ansaloni N, Franchini M (2015) Five variants of a procedure for spatial aggregation of synthetic water demand time series. J Water Supply Res Technol 64(5):629–639

Bennett C, Stewart RA, Beal CD (2013) ANN-based residential water end-use demand forecasting model. Expert Syst Appl 40(4):1014–1023

Buchberger SG, Nadimpalli G (2004) Leak estimation in water distribution systems by statistical analysis of flow readings. J Water Resour Plan Manag 130(4):321–329
Buchberger SG, Wu L (1995) Model for instantaneous residential water demands. J Hydraul Eng 121(3):232–246
Cardell-Oliver R (2013) Water use signature patterns for analyzing household consumption using medium resolution meter data. Water Resour Res 49(12):8589–8599
Cheifetz N, Noumir Z, Samé A, Sandraz AC, Féliers C, Heim V (2017) Modeling and clustering water demand patterns from real-world smart meter data. Drinking Water Engineering and Science 10(2):75–82
Chen J, Boccelli D (2014) Demand forecasting for water distribution systems. Procedia Engineering 70:339–342
Cole G, Stewart RA (2013) Smart meter enabled disaggregation of urban peak water demand: precursor to effective urban water planning. Urban Water J 10(3):174–194
Cominola A, Giuliani M, Piga D, Castelletti A, Rizzoli AE (2015) Benefits and challenges of using smart meters for advancing residential water demand modeling and management: a review. Environ Model Softw 72:198–214
Davidian M, Carroll RJ (1987) Variance function estimation. J Am Stat Assoc 82(400):1079–1091
Firat M, Turan ME, Yurdusev MA (2010) Comparative analysis of neural network techniques for predicting water consumption time series. J Hydrol 384(1–2):46–51
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 Hydroinformatics 18(1):4–22
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 component regression technique. Water Resour Manag 31(1):299–312
House-Peters LA, Chang H (2011) Urban water demand modeling: review of concepts, methods, and organizing principles. Water Resour Res 47(5):W05401
Hutton CJ, Kapelan Z (2015) A probabilistic methodology for quantifying, diagnosing and reducing model structural and predictive errors in short term water demand forecasting. Environ Model Softw 66:87–97
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
Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol Cybern 43(1):59–69
Kottegoda NT, Rosso R (2008) Applied statistics for civil and environmental engineers, 2nd edn. Wiley, UK
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
Madsen H, Thyregod P (2010) Introduction to general and generalized linear models. Chapman & hall/CRC texts in statistical science. CRC Press, Boca Raton
Magini R, Pallavicini I, Guercio R (2008) Spatial and temporal scaling properties of water demand. J Water Resour Plan Manag 134(3):276–284
Magini R, Capannolo F, Ridolfi E, Guercio R (2017) Demand uncertainty in modelling WDS: scaling laws and scenario generation. WIT Trans Ecol Environ 210:735–746
Mamade A, Loureiro D, Covas D, Coelho ST, Amado C (2014) Spatial and temporal forecasting of water consumption at the dma level using extensive measurements. Procedia Engineering 70:1063–1073
McCullagh P, Nelder J (1989) Generalized linear models. Chapman & hall/CRC monographs on statistics and applied probability. CRC Press, Boca Raton
Nasseri M, Moeini A, Tabesh M (2011) Forecasting monthly urban water demand using extended Kalman filter and genetic programming. Expert Syst Appl 38(6):7387–7395
Padulano R, Del Giudice G (2018) A mixed strategy based on Self-Organizing Map for water demand pattern profiling of large-size smart water grid data. Water Resour Manag 32(11):3671–3685
Quevedo J, Puig V, Cembrano G, Blanch J, Aguilar J, Saporta D, Benito G, Hedo M, Molina A (2010) Validation and reconstruction of flow meter data in the Barcelona water distribution network. Control Eng Pract 18(6):640–651
Tricarico C, De Marinis G, Gargano R, Leopardi A (2007) Peak residential water demand. In: Proceedings of the institution of civil engineers–water management, vol 160. Thomas Telford Ltd, pp 115–121

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

Vertommen I, Magini R, da Conceição Cunha M (2015) Scaling water consumption statistics. J Water Resour Plan Manag 141(5):04014072

Wa'el AH, Memon FA, Savic DA (2016) Assessing and modelling the influence of household characteristics on per capita water consumption. Water Resour Manag 30(9):2931–2955

Wong JS, Zhang Q, Chen YD (2010) Statistical modeling of daily urban water consumption in Hong Kong: Trend, changing patterns, and forecast. Water Resour Res 46(3):W03506

Zhou S, McMahon T, Walton A, Lewis J (2002) Forecasting operational demand for an urban water supply zone. J Hydrol 259(1–4):189–202