Water Resour Manage (2016) 30:3443–3457
DOI 10.1007/s11269-016-1361-3

A Comprehensive Approach for Spatial and Temporal Water Demand Profiling to Improve Management in Network Areas

Dália Loureiro[1] · **Aisha Mamade**[2] · **Marta Cabral**[1] · **Conceição Amado**[3] · **Dídia Covas**[2]

Received: 5 June 2015 / Accepted: 17 May 2016 /
Published online: 6 June 2016

Abstract The aim of this paper is to present a comprehensive approach for spatial and temporal demand profiling in water distribution systems. Multiple linear regression models for estimating network design parameters and decision trees for predicting daily demand patterns are presented. Proposed approach is a four-step procedure: data collection, data processing, data characterization, and spatial and temporal demand profiling. Continuous flow measurements and infrastructure and billing data were collected from a large set of water network areas and combined with census data. Main results indicate that family structures (i.e., families with elderly or adolescents), individuals' mobility (i.e., people employed in the tertiary sector and university graduates) and public consumption (i.e., public spaces' irrigation) are key-variables to profile water demand. Profiling models are of the utmost importance to describe water demand in areas with no monitoring but with similar socio-demographic characteristics to the ones analyzed, to improve network operation and to support network planning and design in new areas. Obtained models have been tested for new areas, showing good prediction performances.

Keywords Demand profiling · Network area · Regression analysis · Data reduction · Water consumption · Decision trees

✉ Dália Loureiro
dloureiro@lnec.pt

[1] Urban Water Unit, National Civil Engineering Laboratory, Avenida Brasil 101, 1700-066 Lisbon, Portugal

[2] CERIS, Instituto Superior Técnico, Universidade de Lisboa, Avenida Rovisco Pais 1, 1049-001 Lisbon, Portugal

[3] CEMAT, Instituto Superior Técnico, Universidade de Lisboa, Avenida Rovisco Pais 1, 1049-001 Lisbon, Portugal

1 Introduction

The approach used to classify patterns based on large amounts of data is known as profiling (Wright 2009). Profiling can be applied in multiple domains and for a variety of purposes. For instance, in the electricity sector, network managers need to assess the type of demand to balance electricity between generation, transmission and distribution, to support their long-term planning (Espinoza et al. 2005). In the water sector, managers need to have reliable predictions of daily average consumption, peak factors and minimum night consumption for the operation of water distribution systems. Profiling water consumption is challenging given the nature and quality of available data (i.e., different sources and with different temporal and spatial resolutions), the numerous consumption drivers and the horizons and spatial scales involved (Cabral et al. 2014; Donkor et al. 2012).

Water consumption in network areas is mostly influenced by socio-demographic, billing and climate factors (Arbués et al. 2010; Browne et al. 2013; Parker and Wilby 2013). Higher consumption is typically associated with wealthier individuals living in newer and larger households with outdoor uses, such as irrigation and swimming pools (Beal and Stewart 2011). Households with more elements tend to have lower water per capita consumption and households with elderly have patterns of more frugal water consumption (March et al. 2010). In terms of billing and pricing, water is price-inelastic, but the outdoor uses component is sensitive to price rising (Grafton et al. 2011; Tanverakul and Lee 2012). Regarding climate, temperature and precipitation also affect water uses (Polebitski and Palmer 2010).

Until now, demand profiling has been developed at the household level or at the city or region level (Al-Zahrani and Abo-Monasar 2015; Hollermann et al. 2010; Idowu et al. 2012; Polycarpou and Zachariadis 2013; Scheepers and Jacobs 2014), taking into consideration a limited number of influential factors. Hardly any research has been carried out at the census or network area level (Alcocer-Yamanaka and Tzatchkov 2012; Fontdecaba et al. 2012), important for many water planning decisions. Additional, very prediction models incorporate seasonal or daily consumption scenarios, despite the differences in consumption between winter and summer (Polebitski and Palmer 2010) or weekdays and weekends (Alvisi et al. 2007).

The main objective of this paper is to present a comprehensive approach for spatial and temporal demand profiling in network areas, focusing on domestic consumption. Spatial profiling focuses on estimating consumption variables and patterns in network areas without metering but with similar socio-demographic characteristics to the ones analyzed. Temporal profiling focuses on predicting daily and seasonal demand behaviours in a specific area. This approach was explored through the use of extensive data collected from different network areas in Portugal north and south regions. High-resolution flow data (15-min) have been collected during one year, which allowed identifying different seasonal and daily scenarios.

The main contribution of this study is the comprehensive approach for consumption profiling that results in: (i) regression models for estimating design parameters and (ii) daily consumption patterns for different seasonal and daily scenarios. These allow accurate estimations of peaking factors, daily consumption patterns and minimum night consumption, essential for the network operation and management (e.g., water losses control, pumping cost minimization) and network planning, design and rehabilitation.

2 Methodology

The methodology for spatial and temporal consumption profiling involves a four-step procedure – *data collection*, *data processing*, *data characterization*, and *spatial and temporal consumption profiling* – described in the following sections.

2.1 Data Collection

Data collection can be divided in two stages: the first includes collecting data from different network areas and the second collecting census data provided by Statistics Institutes.

In the first stage, flow time series, billed consumption and infrastructure data are collected. Flow data readings from the utilities' SCADA or telemetry systems should be collected for each metered area. The following criteria have been set to select network areas for the analysis:

1. Boundaries of each area where network operation are kept constant along the year.
2. Network areas with annual domestic billed consumption higher than 80 %, to ensure that areas are mainly composed by residential clients.
3. Number of service connections between 150 and 5000, which corresponds to an acceptable network size for operational management (Farley and Trow 2003; Jankovic-Nišic et al. 2004).
4. Service connections with geographic reference. Infrastructure data is available in a geographical information system.
5. High-resolution flow data with a 10 to 15-min time-step and a minimum one-year data record.

In the second stage, socio-demographic data with the smallest territorial division, the "census areas" should be collected. This division corresponds to a homogeneous building and living zones, with ca. 300 households (INE 2012); a network area may include several census areas.

2.2 Data Processing

Flow time series need to be validated, normalized and cleaned. Data validation includes detecting and correcting outliers. Data normalization aims at obtaining data with a regular time step (15 min). Infrastructure and billing data should be standardized and organized to a common database.

A geoprocessing tool to relate infrastructure and billing data from network areas to socio-demographic data organized in census areas should be used. This tool should convert sociodemographic data at the census area level to data at the network area level.

2.3 Data Characterization

This step involves calculating all the variables for demand profiling. Regarding flow time series, consumption variables and daily patterns different seasonal and daily scenario should be considered.

Consumption variables should include peaking factors and average consumption (relevant for pipe design and rehabilitation) and minimum night consumption, and average consumption during minimum night consumption period (relevant for network operation and water losses control).

Daily consumption patterns can be obtained by the average values of flow data measurements at each 15-min in the period of analysis. Dimensionless consumption patterns are obtained by dividing each instantaneous values by the respective daily average. This is useful to compare the daily behaviour of different network areas.

Consumption scenarios can be obtained using hierarchical cluster analyzes (Ward's method and Euclidean distances). The recommended procedure is to first identify of groups of months with a similar behaviour and then, to identify group of weekdays with similar behaviour, within each seasonal scenario following the same approach. Seasonal scenarios are related with changes in outdoor uses throughout the year (e.g., garden watering and swimming pool filling during the summer); daily scenarios are related to water use changes between working days and weekends.

Table 1 presents the consumption variables that should be considered and the respective scenarios (global, seasonal and daily).

Infrastructure variables should include the main characteristics of the pipe network (i.e., material, diameter and installation year) and service connections (i.e., number of service connections and service connection pipe length).

Billing variables should characterize the domestic and major categories of non-domestic consumption (i.e., commerce-industry, collective and public).

Socio-demographic variables should include the four main census categories: building, dwelling, family and individual. Building category refers to building age and number of floors; dwelling category indicates whether the household is used as primary residence, rented or vacant; family category indicates family type and size and individuals' category refers to age, employment and education level. Table 2 shows the 37 infrastructure (8), billing (6) and socio-demographic (23) variables that should be considered for subsequent analyzes.

2.4 Spatial and Temporal Consumption Profiling

Profiling involves setting consumption variables and patterns as dependent variables that will be explained by a combination of socio-demographic, billing and infrastructure independent variables.

Table 1 Analysed time series, consumption variables and scenarios considered

Category of time series	Variable	Consumption scenario
Instantaneous flow values	Instantaneous peaking factor (−)	Global, seasonal, daily
Instantaneous flow values during minimum night consumption period	Average consumption during minimum night consumption period (l/inh hour)	Global, seasonal, daily
Instantaneous minimum flow values during night period	Minimum night consumption value (l/service connection day)	Global, seasonal
Daily flow	Daily peaking factor (−)	Global, seasonal, daily
	Average daily consumption (l/inh day)	
Monthly flow	Monthly peaking factor (−)	Global
	Average monthly consumption l/inh month)	

Table 2 Socio-demographic, infrastructure and billing variables calculated in this study

Category	Sub-category	Variables
Socio-demographic	Building	Buildings until 1970, 1980, 1990, 2000 and up to 2011; Buildings with 1–2 , 3–4 and ≥5 floor
	Dwelling	Residential immobility; Rented dwellings; Vacant dwellings
	Family	Families with adolescents; Families with elderly; Families with unemployed; Small families (1–2 elements); Medium families (3–4 elements); Large families (≥5 elements)
	Individuals	Population above age 65 years; Inactive workers; University graduates; Economic mobility; Active population mobility; Population with 12 years of education
Billing	Domestic	Average domestic consumption (l/inh day); Total domestic consumption (%); Total domestic consumption within each tariff category (%)
	Non-domestic	Total commerce-industry consumption (%); Total collective consumption (%); Total public consumption (%)
Infrastructure	Pipe	Average installation year (year); Average diameter (mm); Total stainless steel pipe length (%); Total grey iron pipe length (%); Total asbestos cement pipe length (%); Unknown material pipe length (%)
	Service-connection	Service connection density (service connection/km); Average service connection pipe length (m)

A model should depend on the fewest number of independent variables (Vandekerckhove et al. 2014). Principal Components Analysis (PCA) should be applied to reduce the number of independent variables into the most significant ones (Jolliffe 2002). Principal Components (PCs) are new orthogonal (uncorrelated) variables given by linear combinations of the original ones that preserve the total variance. Mathematically, PCA is an eigen decomposition of covariance (or correlation) matrix of the original variables. The Kaiser-Meyer-Olkin measure of sampling adequacy (KMO-test) should be used to avoid reducing the variables to an inadequate size (Kaiser 1970). Adequate samples are the ones with KMO values greater than 0.6 and a total explained variance for each category greater than 75 %. After the PCA, Multiple Linear Regression (MLR) should be carried out by setting consumption variables as dependent variables and key-variables as independent variables. For a data set $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$ of n statistical units, the MLR model takes is given by:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_I \qquad i = 1, \ldots, n \tag{1}$$

in which y_i, Dependent or response variable for unit i;

β_P: Regression coefficient related with independent variable p.

x_{ip}: Independent variable p for unit i.

ε_I: Random error at case i.

The regression coefficients $\beta_1 \ldots \beta_p$ represent an increase (positive value) or decrease (negative value) in the expected value of the dependent variable, associated with each independent variable. The expected value of the dependent variable is equal to β_0 when the remaining regression coefficients are null. To evaluate the quality of the results, the standard errors of the estimated regression coefficients should be computed, as well as the adjusted

correlation coefficient R_a^2. This last coefficient is called "adjusted" since it reflects the number of independent variables and the sample size. Additionally, the *p*-value of the overall F-test for the regression model should also be calculated.

For daily consumption patterns, a new cluster analysis with standardized variables needs to be carried out to group areas with similar patterns. A Decision Tree (DT) using CART algorithm and Gini impurity (Breiman et al. 1984) should be calculated to classify consumption patterns.

3 Case-Studies

The methodology was applied to network areas belonging to Portuguese WDS located in two regions: the north region that includes the districts of Oporto (*Por*) and Braga (*Bra*), and the south region incluiing the districts of Lisbon (Lis) and Setúbal (Set). Each area was identified with a code with an abbreviation code followed by the district names (e.g., ADE_Bra refers to an area in Braga district).

Billing and infrastructure data from 86 network areas was collected: 33 areas in the north region and 53 south region. Network length in the studied areas ranged between 4 and 95 km, clients ranged between 740 and 5200 and inhabitants ranged between 2300 and 9300.

Socio-demographic data referring to the last census in Portugal (2011) were obtained through the National Statistics Institute website (www.ine.pt). To convert data at the census area level to the network area level, a geoprocessing tool has been used (Loureiro, 2010; Mamade 2013). This conversion was carried out by weighting census areas according to the number of domestic clients. This weighting option proved to be more accurate than the original weighting method that relied on the Thiessen's polygons of each service connection. A cluster analysis has been applied to highlight main regional differences in terms of the socio-demographic, billing and infrastructure characteristics (Fig. 1).

Regarding billing variables, domestic billed consumption is higher in the north region (Groups 1–2) where families are larger (higher percentage of families with adolescents). Nevertheless, per capita consumption in the north ranges between 50 and 70 l/inh day, whereas in the south (Groups 3–5) it is considerably higher (100–140 l/inh day). This can be due to the higher economic mobility in the south region (higher percentage of workers employed in the tertiary sector), which is typically correlated with higher incomes and may lead to less conservation attitudes towards the use of water. This difference may also be related with the existence of households in the north region that are not connected to the WDS (e.g., households with private wells). The northern region is also characterized by lower temperatures (T) in the summer and much higher precipitation (P) than the south region. This explains the lower public consumption in the northern region.

Regarding socio-demographic and infrastructure variables, the north region has newer buildings, less asbestos cement pipes and a higher proportion of buildings with 1–2 floors, comparatively to the south region. Northern areas also exhibit a higher proportion of families with adolescents, while the southern areas have more elderly population.

Flow time series could not be obtained for all network areas due to insufficient historical data. A total of 17 network areas (5 in the north and 12 in the south) have been used for consumption profiling. Outliers were removed based on the concept of outlier region (Loureiro et al. 2015). For each time series, the minimum consumption during the night period (0 h00-6 h00) was identified and removed from the series, to ensure that neither consumption variables nor daily patterns were influenced by the level of water losses in the network area (Farley and Trow 2003).

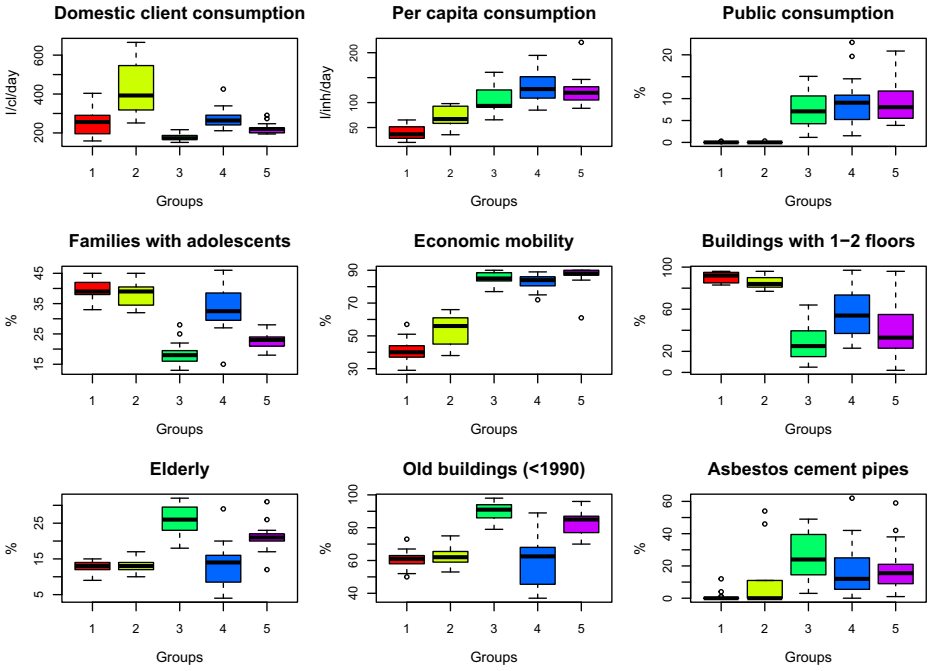

Fig. 1 Cluster analysis revealing the main socio-demographic, billing and infrastructure characteristics of network areas (Group 1–2: North region; Group 3–5: South region)

4 Results from Spatial and Temporal Profiling

4.1 Consumption Scenarios

Daily and seasonal scenarios were identified using cluster analysis. Cluster analysis allowed the identification of two seasons: the winter and the summer seasons. Fig. 2a shows these seasonal scenarios for one network area, in which there is a seasonal average consumption increase and a significant behavioural change. Typically for all the analyzed areas, summer (S) scenarios occur from July to September, whereas winter (W) scenarios occur from November to February.

The next step was to understand daily consumptions behaviours. Results have shown that working days have a different behaviour from Saturdays and Sundays and bank holidays, for both seasonal scenarios. Thus, for spatial and temporal profiling three daily scenarios were analyzed for each seasonal scenario (Fig. 2b).

4.2 Data Reduction

The multiple variables obtained in each category (infrastructure, billing and customers and socio-demography) were reduced to a subset of independent variables using PCA. Since significant regional differences were identified, PCA was separately carried out for the north and south regions. A total of 33 and 53 areas in the north and south regions, respectively, were used in PCA. Table 3 summarizes all the key-variables

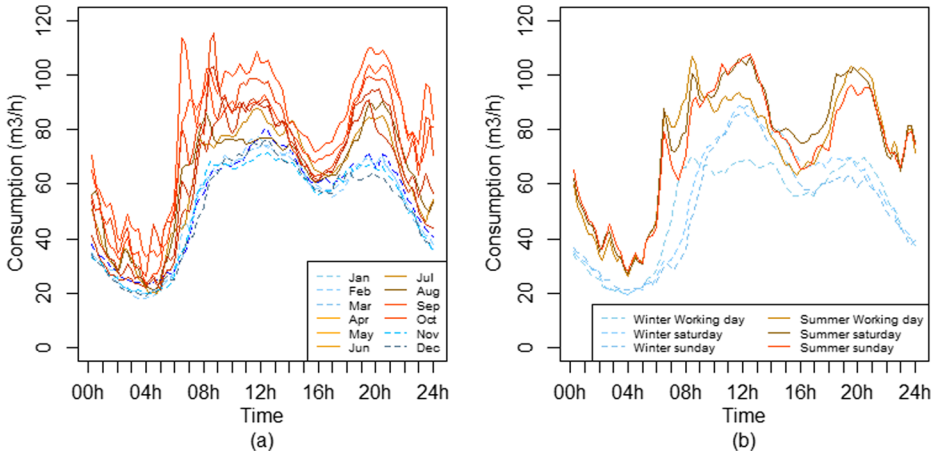

Fig. 2 Scenario exploration for FAR_Set demand patterns: **a** seasonal; **b** daily

considered for DT and MLR analysis and describes the structure of each principal component.

Concerning the socio-demographic category in the south region, the 1st component (PC1: Elderly families) is the most important, as it explains 58.2 % of total variance and shows that families with 1–2 elements and inactive workers or elderly are related (positive loadings), in opposition (negative loadings) to families with 3–4 elements and with adolescents. The 2nd component (PC2: Individuals Mobility) explains 30.7 % of total variance and shows that individuals with higher graduation (university graduates) and working in the tertiary sector (economic mobility) are related, in opposition to individuals with lower education level. For the north region, PCA showed the same components, however, the Individuals Mobility component had a greater importance, explaining 50.6 % of total variance, whereas the Elderly families component explained 26.0 %.

Regarding the infrastructure in the south region, the 1st component (PC1: Pipe material) explains 44.9 % reflecting pipe material, which is independent of pipe size (PC2: Pipe size) that explains 31.0 %. In opposition, for the north region, Pipe size is more important (explains 42.5 %) than Pipe material (explains only 25.3 %).

For billing variables in the south region, PCA was only applied to domestic billed consumption variables as these are independent from non-domestic ones. The only component obtained is PC1: Domestic billed consumption. The same results were obtained for the north region.

In summary, data reduction allowed reducing the 49 initial variables into 8 new variables (5 PCs and 3 variables). A good structure (with high explained variance and KMO) was obtained for both regions and important regional differences were observed.

4.3 Regression Models

A correlation matrix was calculated to analyze which relations between consumption variables (Table 1) and key-variables (Table 3) ought to be explored. After analysing the most significant

Table 3 Key-variables considered for MLR using network data from 53 areas in the south region and 33 areas from the north region

Category	Key-variable	Relevant variable for the South region (loading)	Relevant variable for the North region (loading)
Socio-demography	PC1: Elderly families	Inactive workers (0.92); Elderly (0.94); Families with 1–2 elements (0.95); Families with 3–4 elements (−0.96); Families with adolescents (−0.97)	Elderly (0.90); Families with adolescents (−0.77)
	PC2: Individuals mobility	Economic mobility (0.80); University graduates (0.97); People with 12 years of education (−0.91)	Economic mobility (0.80); University graduates (0.86); Families with >5 elements (− 0.81); Families with 1–2 elements (0.88)
Infrastructure	PC1: Pipe material	Plastic pipes (−0.82); AC pipes (0.83); Service connection density (0.74)	AC pipes (0.93)
	PC2: Pipe size	% Diameter 110–310 (0.77); % Diameter ≤ 110 (−0.78)	Plastic pipes (0.77); % Diameter 110–310 (−0.41); % Diameter ≤ 110 (−0.93); Service connection density (0.70)
Billed consumption	PC1: Domestic Billed consumption	Domestic consumption per inhabitant (−0.61); Domestic Consumption 1st level (0.78); Domestic Consumption 2nd level (0.93); Domestic Consumption 4th level (−0.91)	Domestic consumption per inhabitant (0.87); Domestic Consumption 1st level (−0.78); Domestic Consumption 2nd level (−0.84); Domestic Consumption 4th level (0.93)
	Commerce and industry billed consumption	Commercial and industrial billed consumption category (%)	Commercial and industrial billed consumption category (%)
	Public billed consumption	Public billed consumption category (%)	Public billed consumption category (%)
	Collective billed consumption	Collective billed consumption category (%)	Collective billed consumption category (%)

correlations, a MLR analysis was separately carried out for both regions. Obtained regressions are presented in Table 4.

In the south region, domestic billed consumption is negatively influenced by the *Elderly families* component ($\beta_2 = -25.8$) showing that families with 3–4 elements with adolescents consume more water for domestic uses, which is coherent with the north region results. This variable also negatively relates with the *Pipe size* component ($\beta_2 = -13.6$), meaning that increases with higher pipe diameters (above 110 mm).

In terms of the average consumption per inhabitant, two seasonal scenarios were analyzed: winter and summer. For both scenarios, the average consumption per inhabitant is higher for individuals with higher mobility ($\beta_1 = 34.2$) and monthly consumptions above $25m^3$ ($\beta_2 = -21.3$). Consumption is higher in the summer ($\beta_0 = 220.5$) than in winter ($\beta_0 = 172.1$).

Table 4 Profiling models obtained through MLR

Model	Region	N.º areas	Dependent variable	Explaining component	Regression coefficient	Standard-Deviation	*p*-value	R_a^2
A	North	33	Domestic billed consumption per client (l/cl day)	Constant (β_0)	318.0	15.5	0.0002	0.53
				Elderly families (β_1)	−29.4	16.7		
				Individuals mobility (β_2)	62.5	17.9		
				Pipe material (β_3)	65.5	16.2		
B	South	53	Domestic billed consumption per client [l/cl day]	Constant (β_0)	228.9	5.8	0.0001	0.35
				Elderly families (β_1)	−25.8	6.0		
				Pipe size (β_2)	−13.6	6.0		
C	South	12	Daily peaking factor [−]	Constant (β_0)	1.39	0.04	0.0009	0.81
				Elderly families (β_1)	−0.02	0.03		
				Domestic consumption (β_2)	−0.10	0.03		
				Pipe material (β_3)	0.10	0.04		
D	South	12	Average consumption per inhabitant [l/inh day] – winter	Constant (β_0)	57.3	21.0	0.004	0.72
				Individuals mobility (β_1)	16.5	10.3		
				Domestic consumption (β_2)	20.1	7.7		
				Commerce and industry consumption (β_3)	8.1	1.8		
E	South	12	Average consumption per inhabitant [l/inh day] – summer	Constant (β_0)	220.5	12.1	0.015	0.90
				Individuals mobility (β_1)	34.8	16.6		
				Domestic consumption (β_2)	−55.4	14.7		
F	South	12	Minimum night consumption per service connection [l/sc day] – winter	Constant (β_0)	105.5	109.7	0.003	0.73
				Elderly families (β_1)	74.1	42.7		
				Individuals mobility (β_2)	54.0	46.0		
				Commerce and industry consumption (β_3)	28.6	9.9		

The daily peaking factor increases mostly with monthly consumptions above 25 m^3 ($\beta_2 = -0.10$), plastic pipes ($\beta_3 = 0.10$) and families with 3–4 elements and adolescents ($\beta_1 = -0.02$).

The minimum night consumption is analyzed in the winter scenario, since the average flows are generally lower and leakage becomes more significant. This variable is mainly influenced by socio-demographic characteristics, since the infrastructure is recent and in good condition, with low percentage of asbestos cement (AC) (< 30 %) Thus, the minimum night consumption increases with the *Elderly families* component ($\beta_1 = 74.1$) and the *Individuals' mobility* component ($\beta_2 = 54.0$), as well as with *Commerce-industry consumption* ($\beta_3 = 11.1$).

Domestic billed consumption in the north tends to be higher than in the south. This is explained by the family size: in the north, 66 % of the families have more than 3 elements, while in the south this represents 40 %. In this region, domestic billed consumption relates

positively with *Individuals Mobility* component and negatively with the *Elderly Families*. Tertiary sector Employees typically have higher incomes, leading to higher water consumptions with less conservation attitudes (Beal and Stewart 2011). This regression also indicates that consumption relates positively with *Pipe material*, increasing in network areas where AC pipes predominates.

Results are encouraging and should be explored with a large number of network areas.

4.4 Classification of Daily Consumption Patterns

A cluster analysis was firstly used to group dimensionless patterns obtained in the different areas. This CA considered only the winter period, due to the more homogeneous consumption and only the working days, due to the difference of consumption behaviour between working days and weekends. The instantaneous consumption was characterized by the median and the 10th and the 90th percentiles of dimensionless consumption. A total of 18 areas (out of the initial 21) was used for CA.

Four types of daily consumption patterns were obtained (Fig. 3). Consumption is characterized at six periods: transition (6-7 h and 22-1 h), night (1-6 h), morning (7-10 h), lunch (10-15 h), afternoon (15-19 h) and dinner (19-22 h). Obtained patterns are:

- Type 1: maximum value of the consumption in the morning (2.2), lower consumption at lunch and afternoon and a significant consumption at dinner (1.5);
- Type 2: largest consumption during the day (morning, lunch, afternoon and dinner factor higher than 1.0) and morning and dinner peaks with identical c (1.5);

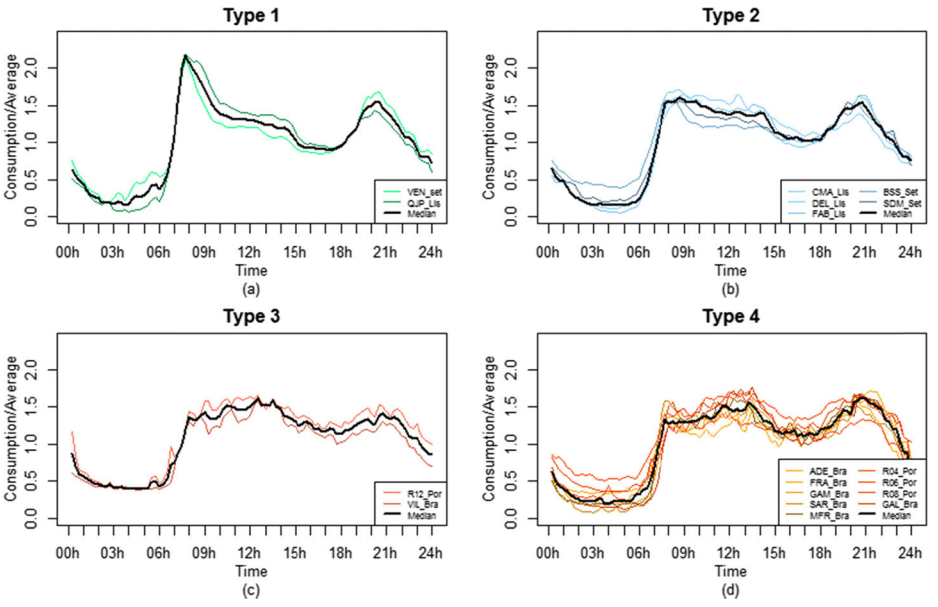

Fig. 3 Daily consumption patterns for working days: **a** Type 1, **b** Type 2, **c** Type 3, and **d** Type 4

- Type 3: higher consumption in the morning and lunch periods (1.5–1.7) and lower consumption in the dinner period (1.4);
- Type 4: largest and identical consumption at lunch and dinner (1.6) and a significant consumption during the morning period (1.4).

Types 1 and 2 correspond to the areas from the south region, wherein the economic and individuals' mobility is higher and the individuals spend most time out. This fact justifies morning and dinner peaks and lower consumption during the day.

Types 3 and 4 correspond to the areas from the north region, wherein the consumption factors are higher throughout the day (morning, lunch, afternoon and dinner period), more similar to consumption patterns in the weekends. These areas present a lower percentage of active population of 47 % (against 68 % from the south), allowing consumption throughout the day.

Decision tree results used to classify the daily demand pattern on working days is presented in Fig. 4. The decision tree was constructed only using two variables to classify the consumption patterns: public billed consumption and individuals' mobility.

4.5 Validation

The regression models have been tested and validated using three new network areas. A relative error given by the absolute difference between the real and the estimated value divided by the estimated value was used. A network area in the north region (VIL_Bra) has been used to validate the Model A (domestic billed consumption). Two network areas in the south region were selected to validate Models C (daily peaking factor) and Model D (per capita consumption in the summer). Table 5 presents the relative errors, showing that the models have a good prediction performance, particularly for the south region, since theses models have a higher R_a^2.

Pattern validation included two steps: decision tree application to classify two new sectors (ALF_Lis and QTE_Lis were classified as Type1); and comparison with the median pattern of the classified pattern. Figure. 5 shows that the real patterns are close to the median of Type 1 pattern.

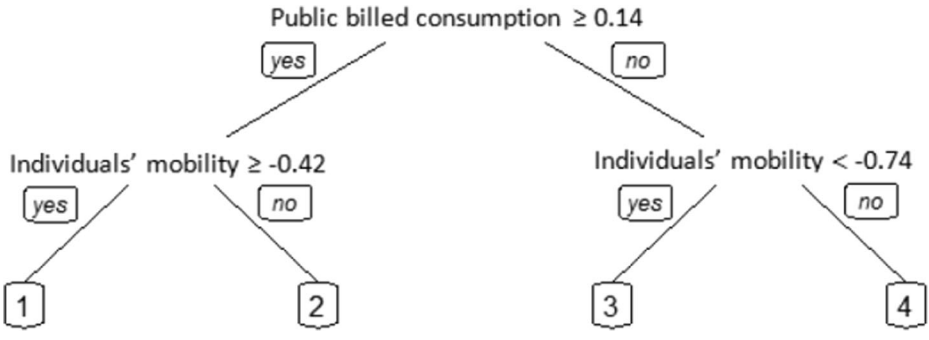

Fig. 4 Decision tree to classify the daily demand pattern for working days

Table 5 Relative errors for the validation models

Model	Network area	Relative error [%]
Model A	VIL_Bra	21
Model C	ALF_Bra	9
	QTE_Lis	7
Model D	ALF_Bra	14
	QTE_Lis	7

5 Concluding Remarks

This research aims at developing a comprehensive approach for spatial and temporal profiling of water consumption variables and patterns in WDS. The approach is applied to 86 network areas considering 49 initial socio-demographic, billing and infrastructure variables. Scenario exploration allowed the identification of seasonal (winter and summer) and daily scenarios (working days, Saturdays and Sundays).

Principal Components Analysis was carried out to define the most relevant socio-demographic, billing and infrastructure variables (reducing the number of variables to 8), followed by Multiple Linear Regression models to estimate average, peaking and minimum consumption. The most important components obtained were socio-demographics and indicate that family structures (i.e., families with elderly or adolescents), individuals' mobility (i.e., people employed in the tertiary sector and university graduates) and public consumption (i.e., public spaces' irrigation) are key-variables to profile water demand.

Four different consumption patterns have been identified, clearly showing that different daily consumption behaviours are mainly associated with different family structures (families with adolescents or elderly).

Results are encouraging and should be explored with a larger number of areas. This research considerably reduces the uncertainty in planning and operation of water distribution systems, thereby improving their efficiency and sustainability.

Acknowledgments The authors would like to thank Rita Almeida from AGS, S. A, Catarina Sousa from Águas do Sado, S.A. and André Pina from SMAS de Oeiras e Amadora water utilities who contributed to this work by providing their data. Authors also thank Alexandre Santos from Alwadi company, for contributing with his expertise in the geoprocessing tool.

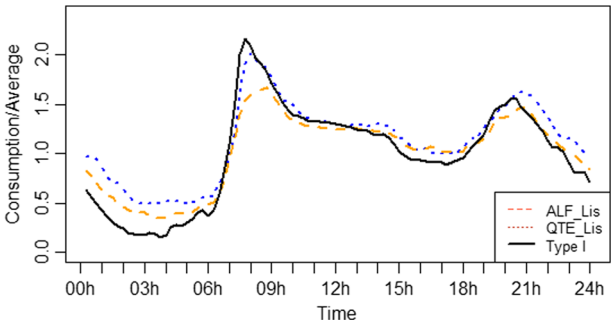

Fig. 5 Pattern validation using ALF_Lis and QTE_Lis

References

Alcocer-Yamanaka VH, Tzatchkov VG (2012) Modeling of drinking water distribution networks using stochastic demand. Water Resour Manag 26:1779–1792

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

Al-Zahrani MA, Abo-Monasar A (2015) Urban residential water demand prediction based on artificial neural networks and time series models. Water Resour Manag 29:3651–3662

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

Beal C, Stewart RA (2011) South East Queensland residential end use study: final report. Griffith University

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA

Browne A, 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 27:1013–1035. doi:10.1007/s11269-012-0117-y

Cabral M, Loureiro D, Mamade A, Covas D (2014) Water demand projection in distribution systems using a novel scenario planning approach. Procedia Engineering 89:950–957

Donkor EA, Mazzuchi TA, Soyer R, Alan Roberson J (2012) Urban water demand forecasting: review of methods and models. J Water Resour Plan Manag 140:146–159. doi:10.1061/(ASCE)WR.1943-5452.0000314

Espinoza M, Joye C, Belmans R, De Moor B (2005) Short-term load forecasting, profile identification, and customer segmentation: a methodology based on periodic time series power systems. IEEE Transactions on 20:1622–1630

Farley M, Trow S (2003)) Losses in water distribution networks. A practitioner's guide to assessment, monitoring and control. IWA Publishing, London

Fontdecaba S, Grima P, Marco L, Rodero L, Sánchez-Espigares J, Solé I, Tort-Martorell X, Demessence D, Martínez De Pablo V, Zubelzu J (2012) A methodology to model water demand based on the identification of homogenous client segments. Application to the city of Barcelona. Water Resour Manag 26:499–516. doi:10.1007/s11269-011-9928-5

Grafton RQ, Ward MB, To H, Kompas T (2011) Determinants of residential water consumption: evidence and analysis from a 10-country household survey. Water Resour Res 47. doi:10.1029/2010WR009685

Hollermann B, Giertz S, Diekkruger B (2010) Benin 2025-balancing future water availability and demand using the WEAP 'Water evaluation and Planning' system. Water Resour Manag 24:3591–3613

Idowu OA, Awomeso JA, Martins O (2012) An evaluation of demand for and supply of potable water in an Urban Centre of Abeokuta and environs, southwestern Nigeria. Water Resour Manag 26:2109–2121

INE (2012) Census 2011: final results - Portugal. Instituto Nacional de Estatística - Statistics Portugal, Lisbon

Jankovic-Nišic B, Maksimovic C, Butler D, Graham NJ (2004) Use of flow meters for managing water supply networks. J Water Resour Plan Manag 130:171–179

Jolliffe I (2002) Principal component analysis. Wiley Online Library

Kaiser HF (1970) A second generation little jiffy. Psychometrika 35:401–415

Loureiro D (2010) Consumption analysis methodologies for the efficient management of water distribution systems (in portuguese). PhD Thesis, Universidade Técnica de Lisboa

Loureiro D, Amado C, Martins A, Vitorino D, Mamade A, Coelho ST (2015) Water distribution systems flow monitoring and anomalous event detection: A practical approach. Urban Water J:1–11. doi:10.1080/1573062X.2014.988733

Mamade A (2013) Profiling consumption patterns using extensive measurements - a spatial and temporal forecasting approach for water distribution systems. Universidade Técnica de Lisboa, MSc Thesis

March H, Perarnau J, Saurí D (2010) Exploring the links between immigration, ageing and domestic water consumption: the case of the metropolitan area of Barcelona. Reg Stud 46:229–244. doi:10.1080/00343404.2010.487859

Parker J, Wilby R (2013) Quantifying household water demand: a review of theory and practice in the UK. Water Resour Manag 27:981–1011. doi:10.1007/s11269-012-0190-2

Polebitski AS, Palmer RN (2010) Seasonal residential water demand forecasting for census tracts. J Water Resour Plan Manag 136:27–36

Polycarpou A, Zachariadis T (2013) An econometric analysis of residential water demand in Cyprus. Water Resour Manag 27:309–317

Scheepers H, Jacobs H (2014) Simulating residential indoor water demand by means of a probability based end-use model. J Water Supply Res Technol AQUA 63(6):476–488

Tanverakul SA, Lee J (2012) Historical review of U.S. residential water demand. In: World Environmental and Water Resources Congress 2012:3122–3136. doi:10.1061/9780784412312.313

Vandekerckhove J, Matzke D, Wagenmakers EJ (2014) Model comparison and the principle of parsimony. In: Busemeyer JR, Townsend Z, Wang J, Eidels A (eds) Oxford handbook of computational and mathematical psychology. Oxford University Press, Oxford In press

Wright D (2009) Profiling the European Citizen: Cross-Disciplinary Perspectives info 11:96–98