Water Resour Manage (2012) 26:499–516
DOI 10.1007/s11269-011-9928-5

# A Methodology to Model Water Demand based on the Identification of Homogenous Client Segments. Application to the City of Barcelona

**Sara Fontdecaba · Pere Grima · Lluís Marco · Lourdes Rodero · José A. Sánchez-Espigares · Ignasi Solé · Xavier Tort-Martorell · Dominique Demessence · Victor Martínez De Pablo · Jordi Zubelzu**

Received: 15 July 2010 / Accepted: 29 September 2011 /
Published online: 20 October 2011


**Abstract** Water management has become a vital concern for both water supply companies and public administrations due to the importance of water for life and current scarcity in many areas. Studies exist that attempt to explain which factors influence water demand. In general, these studies are based on a small sample of consumers and they predict domestic water consumption using ordinary least squares regression models with a small number of socioeconomic variables as predictors, usually: price, population, population density, age, and nationality. We have followed a different approach in two ways; one, in the scope of the study: we have included in the study all consumers of the Barcelona area and as many socioeconomic variables as possible (all the available data from official statistics institutions); and also in the methodology: first, we have segmented clients into homogeneous socioeconomic groups that, as we show later in the Barcelona case, also have homogeneous water consumption habits. This allows for a better understanding of water consumption behaviours and also for better predictions through modeling water consumption in each segment. This is so because the segments' inner variability is smaller than the general one; thus, the models have a smaller residual variance and allow for more accurate forecasts of water consumption. The methodology was applied to the Barcelona metropolitan area, where it was possible to construct a database including both water consumption and socioeconomic information with more than one million observations. Data quality was a primary concern, and thus a careful exploratory data analysis procedure led to a careful treatment of missing observations and to the detection and correction or removal of anomalies. This has resulted in a stable division of the one million water consumers into 6 homogeneous groups and models for each of the groups. Although the methodology has been developed and applied to the Barcelona area, it is general and thus can be applied to any other region or metropolitan area.

S. Fontdecaba · P. Grima · L. Marco · L. Rodero · J. A. Sánchez-Espigares · I. Solé · X. Tort-Martorell (✉)
Department of Statistics and Operational Research, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
e-mail: xavier.tort@upc.edu

D. Demessence · V. Martínez De Pablo · J. Zubelzu
Aigües de Barcelona, AGBAR Group, Barcelona, Spain

**Keywords** Water demand forecast · Segmentation · Cluster analysis · Regression modelling · Scenario planning · SARIMA models

## 1 Introduction

Increasing concerns about the availability of water in adequate quantities and qualities have made more urgent the need to advance towards a sustainable approach to water resources management (Brooks 2006; Butler and Memon 2006; Gleick 2003). In this sense, many international organizations, including the United Nations and the European Union, call for the application of Integrated Water Resources Management (IWRM), which attempts to combine supply and, especially, demand measures (European Commission 2000; ICWE 1992). In particular, governments, water companies and consumers must engage in water conservation practices or otherwise expect growing difficulties in the provision of drinkable water, all probably exacerbated by climate change (Corral-Verdugo et al. 2002; Guy 1996; Postel 1992).

In this context, water demand management—which requires an in-depth knowledge of the behaviour of users in relation to consumption (Dziegielewski 1993; Mazzanti and Montini 2006; Stephenson 1999)– should be a basic component of any water company's strategy (Beecher 1996). Different behaviours, influenced in turn by a variety of factors, explain for instance the high variations in per capita consumption found between different countries, different regions and different cities and even between different neighbourhoods in the same cities.

The study of residential water demand, its characteristics, and its drivers has received substantial attention in the academic and professional literature (Arbués et al. 2003; Babel et al. 2007; Babel and Shinde 2011; Baumann et al. 1998; Duke et al. 2002; Hanke and Mare 1982; Kanakoudis 2002; Opaluch 1982). Different studies and econometric models show that residential water demand is correlated with: income (Arbués and Villanua 2006; Arbués et al. 2003; Baumann et al. 1998; Renzetti 2002); population density (Lavière and Lafrance 1999); age distribution of people in a household (Murdock et al. 1991); religious and cultural characteristics (Smith and Ali 2006); the number of people living in a household (Hamilton 1983; Nauges and Thomas 2003; Renwick and Green 2000; Zhang and Brown 2005); the characteristics of a city or town (Hellegers et al. 2010; Hasse and Nuiss 2007; Kahn 2000; Liu et al. 2003); temperature and rainfall (Griffin and Chang 1991) and the presence of water-saving technology in the form of efficient appliances (USEPA 2005).

With all this in mind, the final objective of the project was to define a methodology for forecasting future water demand, one which benefited from already known facts—that is, from the correlations mentioned above. The idea was to segment clients into homogeneous socioeconomic groups through cluster analysis, then check if these groups were also homogeneous with respect to water consumption (as expected, given the extensive literature mentioned in the previous paragraph) and, finally, use these segments to model water consumption and forecast future demand. Because of the segments' homogeneity we expected the models, one for each segment, to have smaller residual variances and thus allow for more accurate predictions. The project was conducted on behalf of and with full support from Aigües de Barcelona, which was interested in learning about the behaviour of their clients and having accurate water consumption forecasts; thus, we had access to water consumption information for over one million Barcelona area residents. In addition we were able to obtain, through official statistics, socioeconomic information for the same population so we could refine and test the devised methodology.

This paper is organized in the following manner: Section 2 briefly reviews the scope and data used to carry out the study in the area of Barcelona; Section 3 explains the methodology followed in the segmentation and modelling processes and Section 4 highlights and discusses the most relevant results obtained in the Barcelona case; and finally, Section 5 presents some conclusions.

## 2 Case Study and Data Used

Aigües de Barcelona supplies water to nearly 1,570,000 households belonging to 23 municipalities of the Metropolitan Area of Barcelona.

To carry out the study as outlined, two types of data from two different sources were needed: household water consumption, available from Aigües de Barcelona; and sociological variables, obtainable from official statistical institutions. Obviously, in order to detect different customer profiles and group them by similarity, the study required the use of data that was as disaggregated as possible. The consumption data was available at the household level—Aigües de Barcelona meters the consumption of each household—and the minimum territorial division with official statistics data served as the so called "census tracts". The 23 municipalities included in the study contain a total of 2,358 census tracts; they have a population of between 1,500 and 2,000 people and are designed to be homogeneous with respect to population characteristics, economic status, and living conditions. Their spatial size varies widely, depending on population density, and they are fairly stable along time, which allows for statistical comparisons from census to census.

In order to relate census tract socioeconomic data with water consumption data, all supply pipes (distribution pipes connected to several accounts) were assigned to census tracts—a difficult task, since it had to be done "by hand," based on the addresses; however, it was a task that will be extremely useful not only for this but for many other future studies. Only a negligible number of supply pipes were serving more than one census tract or were impossible to locate.

### 2.1 Socioeconomic Data

After an extensive literature review, the most relevant articles for this purpose are listed in the introduction section. Six factors were identified as significant in explaining domestic water consumption: socio-demographic, behavioural, territorial, cultural, climatic and technological. In addition, different variables were defined for each of these factors as possible drivers that influence domestic water demand (Table 1).

Of those factors, "cultural values" and "technology"[1] were not considered in the study, due to a lack of data for the Metropolitan Area of Barcelona. "Price" was also disregarded because it is the same for the whole area. Finally metering, a variable related to price mentioned in several studies, could not be considered for the same reason, in Barcelona all households are metered.

The available data, relative to the rest of the factors, comes from two sources: the census, conducted every 10 years (the most recent available is from 2001); and variables, collected annually by the Instituto Nacional de Estadística (INE), regarding population structure (age,

[1] Partly as a consequence of this study, a survey to gather data on technological variables is scheduled to be conducted.

**Table 1** Factors influencing water demand

| Factors influencing water demand | |
|---|---|
| 1) Price | Economic variables |
| 2) Income | |
| 3) Population and population growth | Socio-demographic variables |
| 4) Household size | |
| 5) Age structure of the population | |
| 6) Cultural values | Behavioural variables |
| 7) Population density | Territorial variables |
| 8) Household size and type | |
| 9) Climate | Climatic variables |
| 10) Technology (water saving appliances in the household) | Technological variables |

sex and nationality). The most recent data available for the latter is from 2006. Variables related to the "Climate" factor were collected from meteorological stations.

All the variables have been collected at the census tract level except for "Income", which is only available at a more aggregate level (the municipality, or, for Barcelona, the so-called Zones de Recerca Petites,[2] which are aggregations of census tracts). To increase the precision of the information and to complement the incomplete variable "Income", we introduced the variable "level of studies" into the database, which is highly correlated with income. The following list presented in Table 2 shows the 27 variables considered in the study, together with a brief description and the units used. They will be referenced in the rest of the paper.

## 2.2 Consumption Data

The main source of information about water consumption comes from the Aigües de Barcelona billing department. Every three months the water consumption of each household account is usually recorded for billing purposes. There were three problems to be solved: estimating the monthly water consumption, getting rid of seasonal effects and aggregating the households that belong to the same census tract. The average daily consumption was estimated over the metered three-month period; then the aggregate consumption of households over the 4 year period was calculated so that its longitudinal evolution (seasonality) could be studied and characterized, and its effects eliminated from each household's monthly consumption data; and finally the estimates of each household were aggregated by supply pipe and census tract.

## 2.3 Exploratory Data Analysis and Data Cleaning

A descriptive analysis of the data gave a preliminary idea of the characteristics of census tracts and of consumer behaviour. At the same time it pointed to some outliers. Specifically, 14 census tracts (0.5% of the 2.358 included in the study) were eliminated due to three different reasons: (I) Lack of sociological information: Newly created census tracts that did not exist in the year 2007. (II) Lack of consumption information: Missing data related to the water consumption of the census tract. (III) Atypical information: Atypical sociological information for a census tract in Barcelona.

[2] Small research zones. They are aggregations of somewhat homogeneous census tracts done for the purposes of sociological studies.

**Table 2** Socioeconomic variables included in the study which could potentially affect water demand

| Variables included in the study | |
|---|---|
| Territorial | |
| Area: Total surface of the census tract ($km^2$) | Density: Population (2006) / Area (inhabitants / $km^2$) |
| Population: Inhabitants registered in the census tract | |
| Sex | |
| %Men: Percentage of men in the census tract | %Women: Percentage of women in the census tract |
| Age Structure | |
| %0-14: *Children*. Percentage of the population between 0 and 14 years | %25–65: *Active population*. Percentage of the population between 25 and 64 years |
| %15-24: *Teenagers*. Percentage of the population between 15 and 24 years | % >65*: Senior population*. Percentage of the population over 64 years |
| Nationality | |
| %Spanish: Percentage of Spanish population | %America: Percentage of American population |
| %Community: Percentage of the population that are Europeans belonging to community countries | %Africa: Percentage of African population |
| %EU Non community: Percentage of the population that are Europeans belonging to non community countries | %Asia: Percentage of Asian population |
| Household | |
| %Principal_Household: Percentage of principal households | %Sec_Or_Empty_Household: Percentage of secondary or empty households (Secondary households are those destined to be occupied only occasionally, i.e. holiday homes. Empty households are those that remain empty without being occupied .) |
| %Small: Percentage of small households (under 60 $m^2$) | %Medium: Percentage of medium households (between 60 $m^2$ and 90 $m^2$) |
| %Big: Percentage of big households (over 90 $m^2$) | Inhabitants_per_Household: Population / # households |
| Income | |
| Income: Per capita Income | |
| %Without_studies: Percentage of population without studies | %Secondary_studies: Percentage of population with secondary studies |
| %Primary_studies: Percentage of population with primary studies | %Higher_studies: Percentage of population with higher studies |
| Population Growth | |
| Increase_Population: The slope of the linear regression of the population from 2003 to 2006. | |

## 3 Methodology

The methodology followed can be divided into two steps. The first one is aimed at obtaining groups of clients. In the Barcelona case, this means groups of census tracts that

have characteristics which are homogenous to the socioeconomic variables that are susceptible to affect water consumption. It is then verified whether or not the groups obtained are also homogeneous in their water consumption. The second step is to find explanatory models and predictive models for the water consumption of each segment. In this section we briefly outline the statistical techniques used in each step.

### 3.1 Segmentation

Cluster analysis (Johnson and Wichern 2002; Lebart, et al. 2006), also called segmentation analysis, seeks to identify homogeneous groups of cases in a population. That is, its objective is to sort cases into groups, or clusters, so that the degree of association is strong between members of the same cluster and weak between members of different clusters. It may reveal associations and structures in data which, though not previously evident, are sensible and useful once discovered. The first step in cluster analysis is the establishment of the similarity or distance matrix. This matrix is a table in which both the rows and columns are the units of analysis and the cell entries are a measure of similarity or distance between any pair of cases; we used the most common one, the Euclidean distance.

For this study, a combination of the two basic families of cluster algorithms (hierarchical and partitional) was used. Hierarchical clustering builds a hierarchy of clusters and is usually represented as a tree diagram (called dendogram) that illustrates the arrangement of the clusters, with individual elements at one end and a single cluster containing every element at the other; cutting the dendogram at a given height gives a clustering at a selected precision. Then, two partitional algorithms were used. First, the *K*-means, which aims to partition all the observations into $k$ clusters, in which each observation belongs to the cluster with the nearest mean. Second, the CART (Classification And Regression Trees), which is a binary recursive partitioning procedure that generates a regression tree.

The steps followed were:

- The socioeconomic variables—omitting consumption—were standardized (subtracted the mean and divided by the standard deviation) in order to give the same weight to all of them, subsequently giving them equal impact on the computation of distances.
- A preliminary classification was obtained using the hierarchical method with the Euclidean distance—because of its interpretability-, and the Ward criterion—in order to achieve a large quantity of groups with a small size. This analysis provided an aggregation tree or dendogram.
- The different classifications (containing different numbers of groups) were used as a starting point for the hierarchical method using the *k*-means algorithm.
- This two-step cluster analysis procedure was repeated including the consumption variables. The same clusters with only minor changes of census tracts appeared, which means that the groups obtained are not only homogeneous both in their structural composition and in their water consumption behavior, but that they are different from each other as well. Therefore, it is reasonable to develop a model for predicting the water consumption of each group.

### 3.2 Modelization

Once the groups were established, the next step was to model water consumption within each group. Because of group homogeneity, it was anticipated that the models obtained would be very good for prediction purposes (the predicted values would have a low

uncertainty). In addition, they would be useful for explaining which variables affect the particular pattern of consumption for each segment.

Two types of models were derived: 1) linear models relating the socioeconomic variables to water consumption useful, as mentioned, for understanding client behavior and rough long-term scenario planning; and 2) predictive time series models (using the SARIMA methodology) for the purposes of accurate short-term predictions.

Linear models (Draper and Smith 1998; Peña 2010) allow explanation of the behavior of a numerical variable (Y) based on values of different variables (Xs). The procedure for finding the linear models is detailed below:

- First, the stepwise algorithm (a semi-automated process of building a model by successively adding or removing variables based on the t-statistics of their estimated coefficients) was used, step by step, to find the best model for explaining water consumption.
- Then, the model found through stepwise regression was fitted and the standardized residuals were obtained to study the goodness of fit and the outliers. In general, a model is considered a good one when the residuals: (a) are normally distributed with mean 0, (b) have a constant variance, (c) are independent. In our study, there were no problems with homoscedasticity or independence. Regarding normality, it was expected to have around 5% outliers (data points poorly explained by the model); that is, points having a residual larger than −2|. In our case, and because of the amount and nature of the data, we moved the bounds from −2| to −4|, and considered outlier points with residuals larger than −4|; these points were studied and removed from the model if no good explanation was found for them. Cook's distance was used to estimate the influence of each data point. Points with Cook's distance larger than 1 were studied and removed if appropriate. In total, six data points were removed, a very small number given the size and type of the study.

### 3.3 Time series

The time series models used for short term prediction purposes were developed using the SARIMA methodology (Box and Reinsel 1994; Peña 2005). This methodology uses the autocorrelation function to measure the relationship between observations within the water consumption series and to derive predictive models. The idea is to describe how any given observation ($x_t$) is related to previous observations ($x_{t-1}$, $x_{t-2}$,…). This model is then used to forecast the future values of the variable. To define the model for each group, we carried out a three-stage procedure:

- Stage 1: Identification. The estimated Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) were calculated and, based on them, the SARIMA model—whose theoretical ACF and PACF most closely resemble the ones estimated from the data—is chosen as the most adequate model.
- Stage 2: Estimation. Via maximum likelihood (ML), we obtained precise estimates of the coefficients of the model chosen in the identification stage.
- Stage 3: Diagnostic Checking. By checking the residuals, we were able to see the suitability of the model fitted for each segment and to detect other effects that influence the behaviour of the series. For example, holiday periods relating to a calendar effect (in our case the Easter period was especially relevant).

Once we have a model for each group, it can be used to make predictions for that group or, by adding the predictions for all the groups, we can obtain a global prediction that is

more precise (with less variability) than if we had obtained it from a global SARIMA model. This is so because the groups' inner variability is smaller than the general variability and, thus, the models have a smaller residual variance that allows for more accurate forecasts of water consumption.

## 4 Results and Discussion

This section shows the results of applying the methodology to the Barcelona case (data explained in Section 2). The results are presented in three blocks: segmentation, modelling and time series. We dedicate more attention to the segmentation and modelling results than to the short term time series forecasting. In the block on segmentation, we explain the groups of customers with homogeneous water consumption habits found in Barcelona while the block on modelling presents the socioeconomic variables that influence water consumption for each group. . More attention is paid to these first two parts because they present new developments: segmentation and the use of segmentation for modelling socioeconomic influences. The third part, on the other hand, merely applies an already well known methodology; the only novelty is its application to homogeneous groups, followed by adding the results so that the total estimate has a smaller variance.

### 4.1 Segmentation of Clients Regarding Water Consumption and Socioeconomic Variables

As stated in Section 3, the point of departure was a hierarchical cluster analysis, conducted using the 29 socioeconomic and water consumption variables. The resulting dendogram is shown in Fig. 1. If an imaginary horizontal line is drawn in the upper part of Fig. 1, and it is slowly moved down along the similarity axes, it will cross a different number of vertical

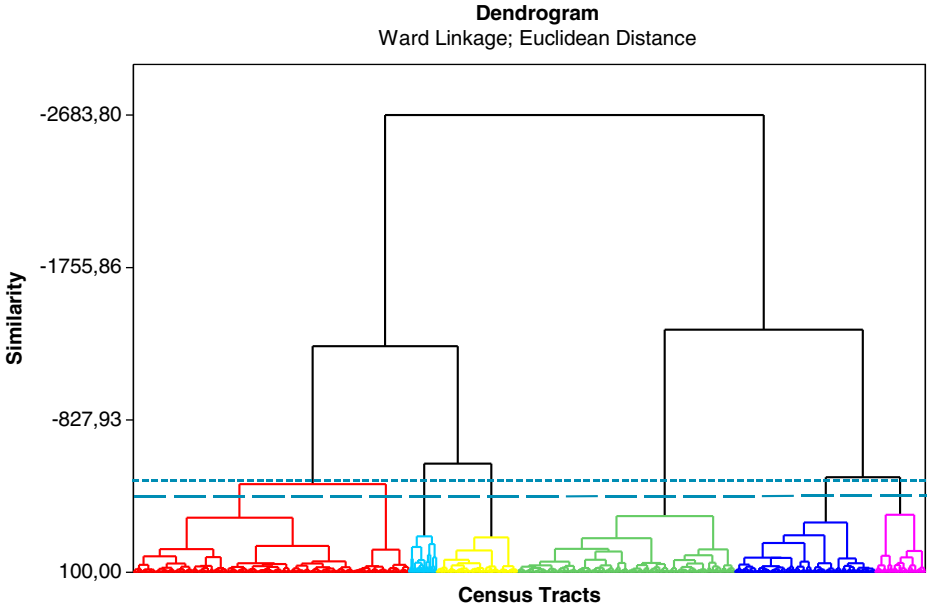


**Fig. 1** Dendogram

lines. The number of vertical lines indicates the number of clusters (groups) into which the data can be classified for a given similarity (shown on the vertical axis). The main doubt arose when deciding between 6 (dotted line) and 7 (hyphenated line) groups. Finally, 6 groups were chosen, both because of their interpretability and because the CART misclassification error rate was very small (slightly above 15%). Furthermore, the six groups were very easy to explain; they "made sense".

Once the 6 groups were established, we characterized them using two different resources. First, using descriptive statistics of the variables to determine the main differences between groups; and second, painting their position on a spatial map to identify the distribution of each group.

The first step is to use descriptive statistics and graphics to get an initial idea of the characteristics of consumers in each group and the relationships between the socioeconomic variables and water consumption.

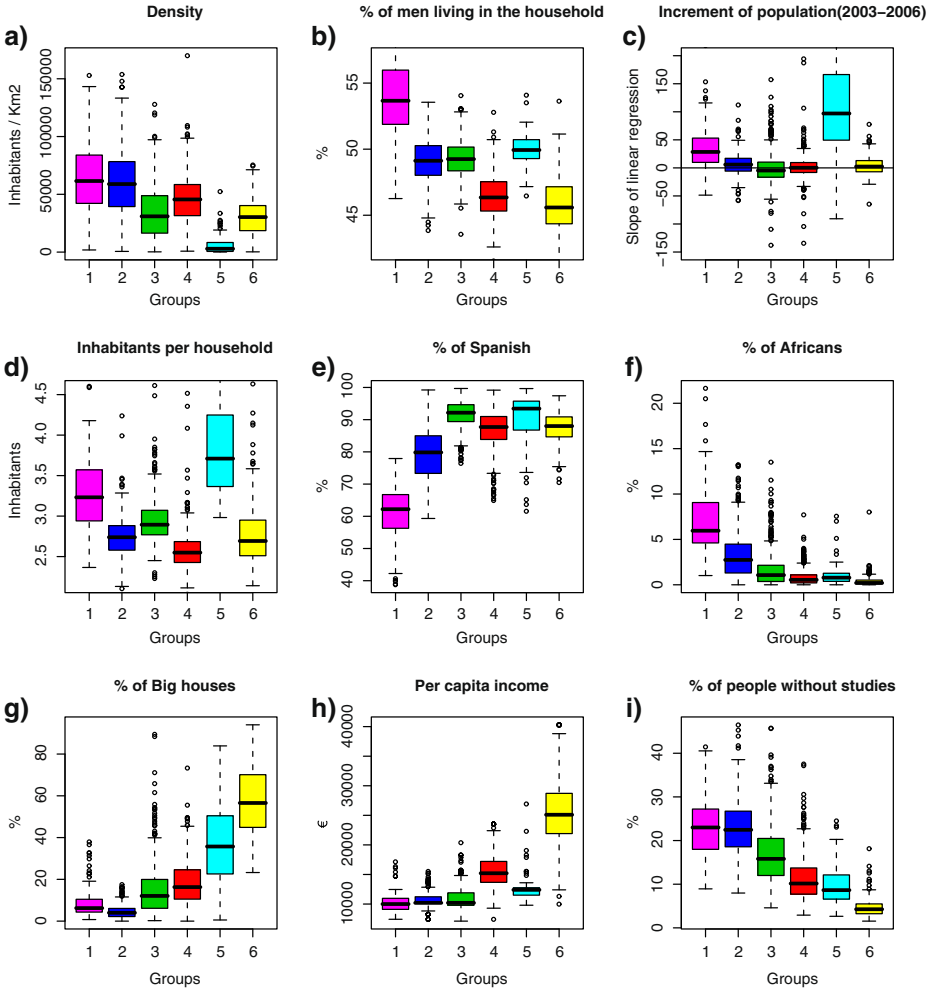


**Fig. 2** *Boxplots* of influential socioeconomic variables (**a**–**i**) stratified by group

Figure 2 shows *Boxplots* stratified by group, showing the socioeconomic variables with the biggest influence on group formations. *Boxplots* are a convenient tool for conveying location and variation information in data sets, particularly for detecting and illustrating location and variation changes between the different groups defined. The box stretches from the lower hinge (defined as the 25th percentile) to the upper hinge (the 75th percentile) and therefore contains the middle half of the scores in the distribution. The median is shown as a line across the box. Therefore 1/4 of the distribution is between this line and the top of the box and 1/4 of the distribution is between this line and the bottom of the box. There are two adjacent values: the largest value below the upper inner fence and the smallest value above the lower inner fence. Outside values, usually considered possible outliers, are indicated by small circles.

The dot plots of Figs. 2 and 3, together with other graphic and data exploratory techniques, allow an interpretation of the socioeconomic characteristics of the groups found. Table 3, below, summarizes the main socioeconomic and water consumption traits of each group.

In our case, since groups are tied to a geographical zone, it is also very instructive to paint them on a map. Figure 4 shows a map of the Barcelona Metropolitan Area with the six groups in different colors. For those familiar with Barcelona's development history and economic structure, it will be easy to recognize the soundness of the clusters found.

Some general trends can be derived from the group characterization:

- As can be seen on the map, groups are distributed roughly in circles from the center. As mentioned previously, it represents a very reasonable pattern, given the history and physical and socioeconomic structure of Barcelona.
- The middle class groups are the largest in the population (which seems quite logical).
- Groups with lower per capita income are the ones with more immigration and less water consumption.
- Groups with higher per capita income have higher water consumption.
- Group number 5, mostly composed of families with a high level of studies, is quite peculiar. It has very high water consumption (gardens) and the highest decrement, but not a very high per capita income (young people starting their career).
- The higher decrements in water consumption come from the groups with higher consumption.

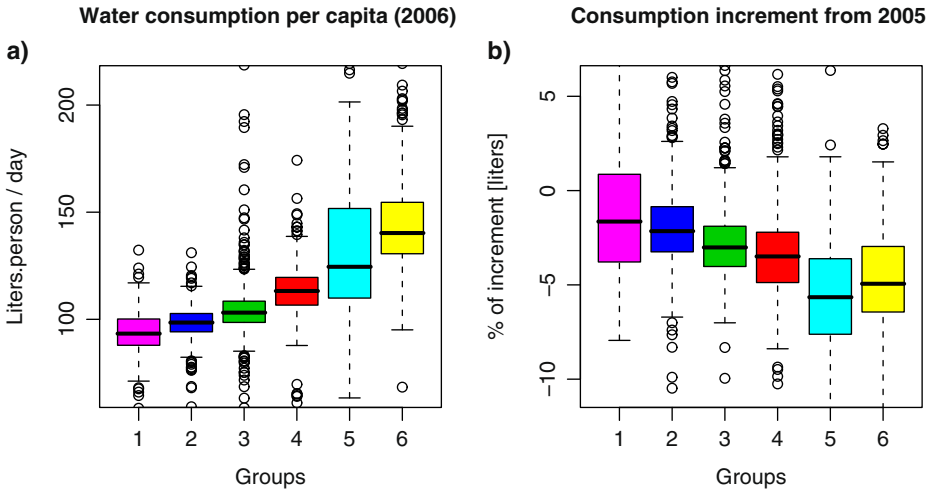


**Fig. 3** *Boxplots* of water consumption (**a**) and consumption increment (**b**), stratified by group

**Table 3** Summary of the socioeconomic characteristics and water consumption per group

| | |
|---|---|
| Group 1. Low income, 40% immigration<br>118 census tracts<br>•The lowest water consumption (93 l pers. /day) and the lowest decrement.<br>•The lowest per capita income<br>•The highest percentage of people without studies or with primary studies.<br>•The highest population density<br>•The highest percentage of immigration (in order, Asians, Americans and Africans)<br>•The highest percentage of men.<br>•A significant population increment.<br>•90% small and middle-sized houses. | Group 2. Low income, 20% immigration<br>411 census tracts<br>•Low water consumption (97 l pers./day) and low decrement.<br>•Low rate of secondary or empty houses<br>•High population density<br>•20% immigration (mainly Americans)<br>•96% small and middle houses. |
| Group 3. Lower middle class<br>634 census tracts<br>•Middle water consumption (105 l pers./day) and decrement in the consumption.<br>•92% Spanish.<br>•The only group with a decrement in population<br>•85% main houses (more than half mid- size) | Group 4. Upper middle class<br>816 census tracts<br>•Middle water consumption (113 l pers./day)<br>•Quite high per capita income<br>•High percentage of elderly people.<br>•More women than men.<br>•40% with secondary or higher studies and only 11% without studies |
| Group 5. Young families<br>73 census tracts<br>•High water consumption (140 l pers./day)<br>•The highest water consumption decrement.<br>•Very low population density and the highest population increment (expanding areas)<br>•The highest percentage of children between 0 and 14 years.<br>•The highest amount of people per household: 4.2.<br>•Low percentage of people without studies.<br>•Mainly middle size and big houses. | Group 6. Wealthy<br>292 census tracts<br>•The highest water consumption (150 l pers./day)<br>•The second highest water consumption decrement.<br>•The highest per capita income<br>•High percentage of elderly people<br>•The highest percentage of women.<br>•Mainly big houses (58%)<br>•The highest percentage of secondary or empty houses<br>•The highest percentage of people with secondary and higher studies.<br>•Almost no increment in population |

## 4.2 Modelling

As referred to in Section 3.2, linear models make it possible to explain the behaviour of a numerical variable (Y), called the dependent variable, based on values of different variables (X), called the independent or explanatory variables. In our case, the dependent variable was the water consumption per capita in 2006–2007; and the explanatory variables, considered at the census tract level, were the 27 socioeconomic variables used to define the segmentation analysis explained in Table 2, plus information about the type of contracts.

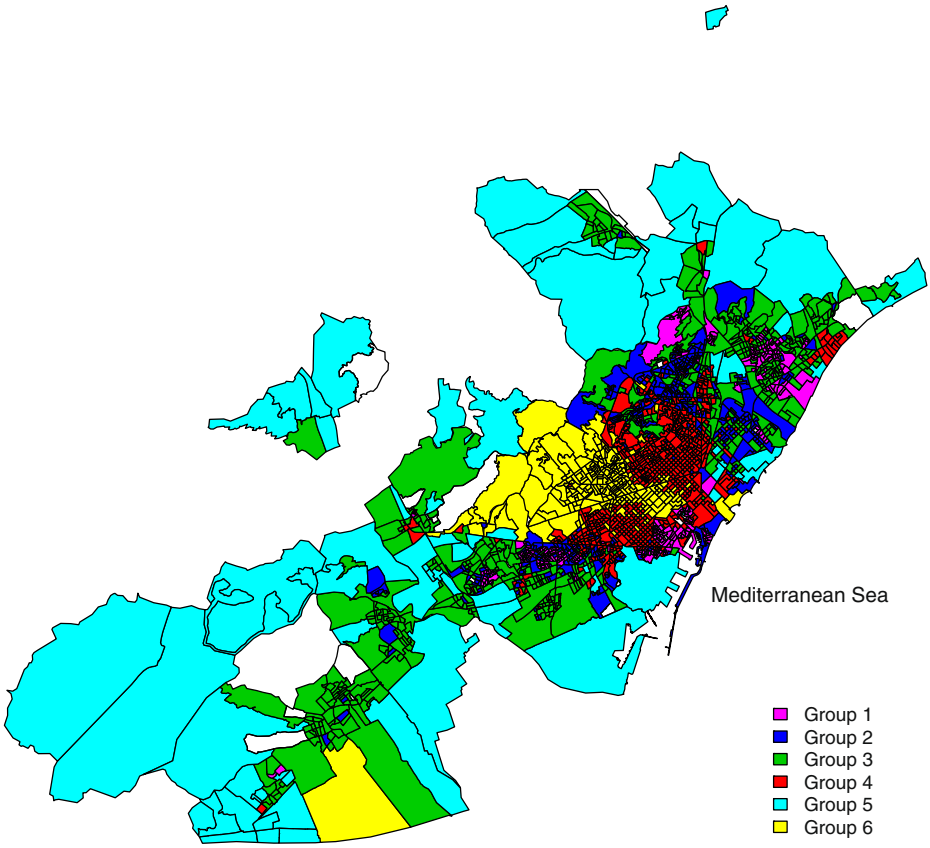


**Fig. 4** Geographical distribution of groups in the Barcelona Metropolitan Area

We have created six variables, defined as percentages of customers having a type of contract.[3]

A linear model was fitted for each of the six groups of consumers identified in the segmentation stage. Thus, for each group a model relating water consumption with socioeconomic variables was obtained. The models are presented in a table form (Table 4) to allow for an easy comparison between groups. The significant variables appear with the actual value of the equation coefficient, the non-significant ones have an asterisk. Thus, the model for Group 1, reproduced from the table, is:

$$\begin{aligned}\text{Water consumption} = {} & 121,55 - 0,0000696 \cdot \text{Density} + 42,95 \cdot \%15 - 24 + 68,08 \cdot \%\text{Community} + \\ & + 16,431 \cdot \%\text{America} - 81,28 \cdot \%\text{Africa} - 11,743 \cdot \%\text{Principal\_Household} - 6,477 \cdot \%\text{Small} + \\ & + 16,912 \cdot \%\text{Big} - 7,9821 \cdot \text{InhabitantsperHousehold} + 0,000218 \cdot \text{Income} + \\ & + 16,991 \cdot \%\text{Secondary\_studies} + 42,379 \cdot \%\text{Higher\_studies} + 1,983 \cdot \%\text{B}\end{aligned}$$

It is important to point out that the socioeconomic variables that do not appear in Table 4 are the variables considered as the base level; thus, the coefficient measures the difference in respect to the base level, which is the usual approach when working with compositional

[3] The different contract types reflect, very broadly, the number of water points in the household.

**Table 4** Regression coefficients for the models

| | Model Group 1 | Model Group 2 | Model Group 3 | Model Group 4 | Model Group 5 | Model Group 6 |
|---|---|---|---|---|---|---|
| Constant | 121,55 | 113,64 | 120,81 | 117,89 | 122.53 | 537,49 |
| Density | −0.0000696 | −0,0000696 | −0,0000696 | −0,000649 | −0,0000696 | −0,0000696 |
| Area | * | * | * | * | * | * |
| %Women | * | * | * | * | * | * |
| %15–24 | 42,95 | 42,95 | 42,95 | 42,95 | 42,95 | 42,95 |
| %25–64 | * | * | * | * | * | * |
| %>65 | * | * | * | 152.87 | −130.55 | * |
| %Spanish | * | * | * | * | 38.339 | −59.07 |
| %Community | 68,08 | 68,08 | −15,02 | 972.76 | 68.08 | 68.08 |
| %EU Non community | * | * | * | 375,8 | * | * |
| %America | 16,431 | 16,431 | 16,431 | −316.069 | 16.431 | 16.431 |
| %Africa | −81,28 | −81,28 | 11,94 | −81,28 | −17.06 | −81.28 |
| %Principal_Household | −11,743 | −11,743 | −11,743 | −11,743 | −11,743 | −34.673 |
| %Small | −6,477 | −6,477 | −6,477 | −45,627 | −6,477 | −6.477 |
| %Big | 16,912 | 16,912 | 16,912 | 16,912 | 16.912 | 35.716 |
| Inhabitants_per_Household | −7,9821 | −1,1421 | −7,9821 | 0,3729 | −7,9821 | −7,9821 |
| Income | 0,000218 | −0.000793 | 0,000218 | −0,001956 | 0,000218 | 0,000218 |
| %Primary_studies | * | * | * | * | * | * |
| %Secondary_studies | 16,991 | 16,991 | 16,991 | 16,991 | 16,991 | 16,991 |
| %Higher_studies | 42,379 | 42,379 | 42,379 | 42,379 | −65.411 | −17.511 |
| %A | * | * | * | * | * | −181.27 |
| %B | 1,983 | 1,983 | 1,983 | 1,983 | 1,983 | −331.577 |
| %C | * | * | * | 21,313 | * | −354.31 |
| %D | * | * | 20,09 | * | * | −339.88 |
| %E | * | * | * | * | * | −337.06 |
| | $R^2$=42% | $R^2$=23% | $R^2$=20 | $R^2$=88% | $R^2$=40% | $R^2$=73% |

data. Several reduced models with approximately the same $R^2$ were found, so that the final choice was based on knowledge of the data and parsimony of the model.

$R^2$ measures the percentage of variability of the per capita water consumption as explained by the variables considered. The explained variability ranges from 20% to 88%, depending on the characteristics and homogeneity of the group. A lower value of $R^2$ does not mean a poor adjustment of the estimated model (in fact, all models are the "best" ones for the available data); it means that the explanatory variables included in the study can only explain a small part of the water consumption. Notice that groups 4 and 6 are very homogeneous and well explained by the model, while groups 2 and 3 have a low $R^2$, meaning that the variability within these groups is high and only partially explained by the variables in the models.

The interpretation and use of these linear models is not straightforward, and has to be done carefully due to the high correlation between the explicative variables and, therefore, between the regression coefficients.

A first and simple use of the models is for pointing out the relative importance of each variable in explaining water consumption and also in estimating water consumption when small changes of the explanatory variables are considered.

A more sophisticated use is scenario planning. That is, to use them in order to have an idea of changes in water consumption when the socioeconomic characteristics change more drastically; for example, when a specific group experiences an increase in per capita income. Because the socioeconomic variables are not independent, it won't be realistic to set a scenario increasing the per capita income and decreasing the level of studies. To help set realistic scenarios, Principal Component Analysis (PCA) is used. PCA is a multivariate statistical technique (Johnson and Wichern 2002) that reduces the dimensionality of a database and shows how variables are related.

Basically, the technique creates new variables (called components) that are linear combinations of the original variables. The components are ranked, and usually just a few of them capture a large amount of variability from the database. That is, only a few components (say, the first two components) have almost the same information as the whole database.

A common representation is the scatter plot showing the weights of the original variables in the first two components of a PCA. In such a graph, the longer the projection of the line for a variable into an axis, the more important the contribution of that variable is for what that axis represents. Figure 5 shows the graph for the principal component analysis made with data from all consumers. Although we do not aim to interpret this graph from a socioeconomic point of view, notice that the horizontal axis contains variables related to economic data (income, level of studies), whereas the vertical axis contains variables related to immigration.

The way to use the PCA graph for scenario planning is the following:

- Variables that are close in the graph should be changed together and in the same direction. Example: Income, % of Secondary_studies and % of Superior_studies.
- Variables that are opposite (at 180º, more or less) should be changed together, but in the opposite direction (if one is increased the other should be decreased). Example: % of men and women.
- Variables that are at 90º are basically independent and thus, can be changed independently. Example: the % of community immigrants (%Community) and % of men and women.

As we have just seen, PCA is a big help in planning scenarios, but still it is very important to use sociological knowledge and common sense. In addition it must be kept in mind that the models reflect the relationship between the socioeconomic variables and water consumption, but not necessarily a cause and effect relationship.

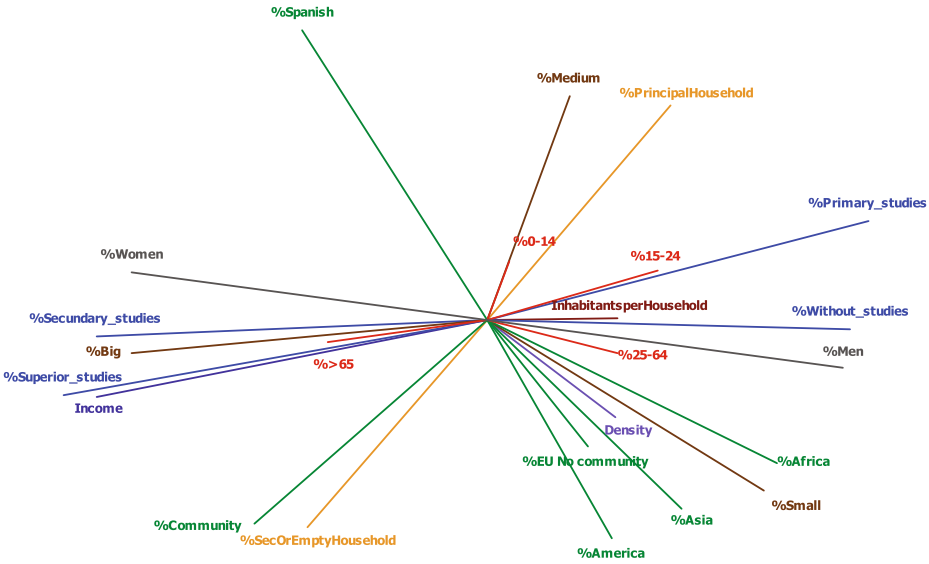


**Fig. 5** Socioeconomic map from a principal component analysis (PCA)

### 4.3 Time Series

As stated earlier, for short term forecasting a time series SARIMA model was fitted to each group's overall consumption. This kind of model is especially suited to short-term forecasting because, in general, it relies heavily on the recent past. In our case, we worked with monthly data. It is reasonable to forecast 6–8 months ahead, up to a maximum of 12. It is also worth noting that these predictions are basically based on past consumption, they don't use socioeconomic data; in our case we forced the models to take into account meteorological data: temperature and rainfall.

The first step was to characterize the profile of monthly consumption—expressed in liters per inhabitant and day—for each group from June, 2003 to December, 2007. We found that the profiles behaved in a similar way with a strong seasonal component, but with different levels of consumption (Fig. 6). Then the SARIMA models were fitted (according to the maximum likelihood criterion) taking into account the calendar and weather effects as they might have an influence.

An example of a calendar effect is the Easter effect. It has to do with the fact that the month of the year that Easter falls in, April or March, has a lower than expected water consumption in the Barcelona area because people leave the city for a few days on vacation. Another calendar effect is due to the fact that the data vary according to the number of times that each day of the week occurs in each month (four or five times); many people also leave the city on weekends. To take into account the impact of climate variables, we imposed that the effects of temperature and rainfall appear in the model.

Then we used Box & Jenkins methodology (Box and Reinsel 1994) to obtain and validate a model for each group. The details of the SARIMA models obtained are omitted since they are not of general interest and the methodology used is well known and has been used in other occasions for water consumption prediction (Molino et al. 1996).

Obviously, from the prediction of consumption for each group, the global monthly water consumption for the Barcelona area can be predicted. The procedure is simple, it has two

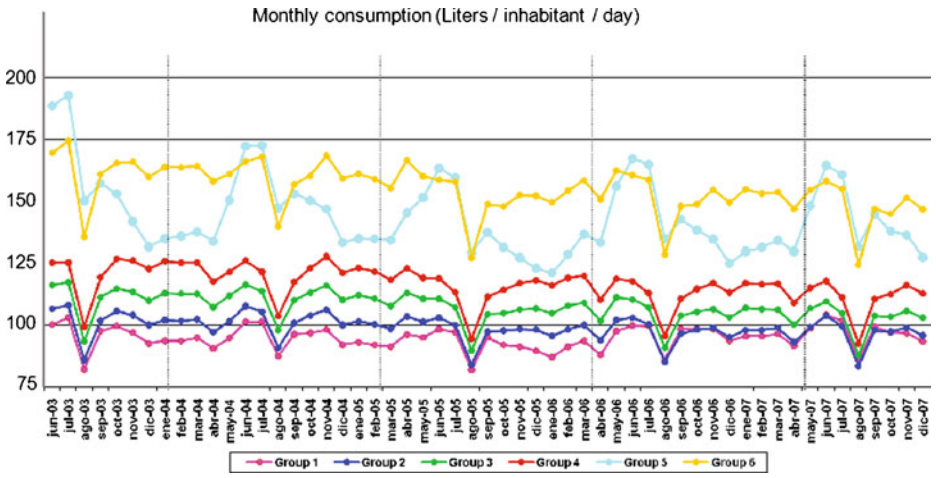


**Fig. 6** Monthly consumption profiles (liter/inhabitant/day) by group from June, 2003 to July, 2007

steps: first, predict the monthly water consumption for each segment; and second, add the results.

What is interesting to note is that the forecast obtained using this method is more reliable and has narrower confidence bands than the one that would have been obtained by modelling the aggregated consumption data directly. As explained in section 3.3, this is because it has a smaller variance. Table 5 shows a comparison of the forecasts for 2007 using the method based on the groups and the method modelling the aggregated consumption.

**Table 5** Comparison of forecast using the aggregated models (based on groups) and global consumption

| Forecast using segmentation (sum of groups forecasts) | | | | Forecast using the aggregated (global) consumption | | | |
|---|---|---|---|---|---|---|---|
| Total Consumption(m3) | | | | Total Consumption(m3) | | | |
| Forecast | Standard Deviation | Lower Limit (95%) | Upper Limit (95%) | Forecast | Standard Deviation | Lower Limit (95%) | Upper Limit (95%) |
| 9,904,222 | 99,405 | 9,705,412 | 10,103,032 | 9,899,908 | 211,577 | 9,476,753 | 10,323,063 |
| 8,872,175 | 99,128 | 8,673,919 | 9,070,431 | 8,860,713 | 210,443 | 8,439,828 | 9,281,599 |
| 9,884,941 | 119,199 | 9,646,544 | 10,123,338 | 9,870,851 | 252,594 | 9,365,662 | 10,376,039 |
| 8,969,167 | 123,825 | 8,721,517 | 9,216,816 | 8,945,840 | 262,048 | 8,421,745 | 9,469,935 |
| 10,051,117 | 136,144 | 9,778,829 | 10,323,406 | 10,028,858 | 287,824 | 9,453,210 | 10,604,506 |
| 9,761,781 | 139,229 | 9,483,323 | 10,040,240 | 9,729,310 | 294,108 | 9,141,094 | 10,317,525 |
| 10,015,510 | 151,202 | 9,713,105 | 10,317,914 | 9,991,192 | 319,189 | 9,352,815 | 10,629,570 |
| 8,146,244 | 158,195 | 7,829,854 | 8,462,633 | 8,119,114 | 333,768 | 7,451,579 | 8,786,649 |
| 9,204,700 | 159,572 | 8,885,555 | 9,523,844 | 9,181,362 | 336,519 | 8,508,325 | 9,854,399 |
| 9,596,037 | 171,326 | 9,253,385 | 9,938,688 | 9,557,126 | 361,164 | 8,834,797 | 10,279,454 |
| 9,512,924 | 171,801 | 9,169,322 | 9,856,525 | 9,477,400 | 362,043 | 8,753,314 | 10,201,485 |
| 9,432,299 | 183,520 | 9,065,260 | 9,799,338 | 9,397,845 | 386,624 | 8,624,596 | 10,171,094 |

## 5 Concluding Remarks

The paper presents a methodology to study, understand, model and forecast water consumption based on socioeconomic data available from official statistics. The statistical methodology used is based on cluster analysis to identify groups with similar consumption habits and socioeconomic status and then on regression analysis and time series SARIMA models to predict and understand water consumption.

The paper confirms, in the Barcelona case, the initial idea that there is a relationship between water consumption habits and socioeconomic characteristics and that this relationship can be used to detect groups of clients with similar consumption patterns and then use these groups for different purposes: a more accurate short term prediction of water consumption, long term scenario planning and a better understanding of consumer habits. The benefits of having the customers segmented into homogeneous groups does not stop here; another important benefit is the ability to take samples of clients in a stratified way—sampling within each cluster—which allows for more reliable results with small sample sizes.

The results from the Barcelona Metropolitan area are the six clusters that were found, used throughout the article, to show the methodology, the difficulties encountered and the results obtained.

The proposed methodology can be used in other areas with only small adaptations. The only requirements are: having socioeconomic data on a census tract level, water consumption at the consumer (or other disaggregate) level and being able to situate the consumers in census tracts. While writing this article we have been in contact with other water companies, which belong to the AGBAR group, in order to apply the methodology.

**Acknowledgements** The authors are grateful to Montserrat Termes from CETAQUA for very useful comments and suggestions during the preparation of this manuscript.

The authors are grateful to R + I Alliance for the financial support that made it possible to develop this project.

## References

Arbués F, Villanua I (2006) Potential for pricing policies in water resource management: Estimation of urban residential water demand in Zaragoza, Spain. Urban Studies 43:2421–2442

Arbués F, García-Valiñas M, Martínez-Espiñeira R (2003) Estimation of residential water demand: a state-of-the-art review. J Socio-Economics 32:81–102

Babel MS, Shinde V (2011) Identifying prominent explanatory variables for water demand prediction using artificial neural networks: a case study of bangkok. Water Resour Manag 25(6):1653–1676. doi:10.1007/s11269-010-9766-x

Babel MS, Das Gupta A, Pradhan P (2007) A multivariate econometric approach for domestic water demand modeling: An application to Kathmandu, Nepal. Water Resour Manag 21(3):573–589. doi:10.1007/s11269-006-9030-6

Baumann DD, Boland J, Hanemann WM (1998) Urban water demand management and planning. McGraw-Hill, New York

Beecher J (1996) Integrated resources planning for water utilities. Water Resources Update, 104

Box, Jenkins and Reinsel (1994) Time series analysis: forecasting and control, 3rd edn, Prentice-Hall, pp 135–168

Brooks DB (2006) An operational definition of water demand management. Water Resources Development 22:521–528

Butler D, Memon F (2006) Water demand management. International Water Association Publishing (IWAP), London

Corral-Verdugo V, Frías-Armenta M, Pérez-Urias F, Orduña-Cabrera V, Espinoza-Gallego N (2002) Residential water consumption, motivation for conserving water and the continuing tragedy of the commons. Environ Manag 30:527–535

Draper N and Smith, W (1998) Multiple regression: special topics. In: Wiley(ed). Applied regression analysis, 3 rd edn, pp 217–234.

Duke JM, Ehemann RW, Mackenzie J (2002) The distributional effects of water quantity management strategies: a spatial analysis. Rev Reg Stud 32(1):19–35

Dziegielewski B (1993) Management of Water Demand: Unresolved Issues. J Water Resour Update 114:1–7

European Commission (2000) EU Water Framework Directive. Directive 2000/60/EC.

Gleick PH (2003) Water use. Annu Rev Environ Resour 28:275–314

Griffin RC, Chang C (1991) Seasonality in community water demand. West J Agric Econ 16(2):207–217

Guy S (1996) Managing water stress: the logic of demand side infrastructure planning. J Environ Plan Manag 39:123–130

Hamilton L (1983) Saving water: a causal model of household conservation. Sociological Perspectives, núm 26:355–374

Hanke S, de Mare L (1982) Residential water demand: a pooled, time series, cross section study of Malmö, Sweden. J Am Water Resour As 18(4):621–626. doi:10.1111/j.1752-1688.1982.tb00044.x

Hasse D, Nuiss H (2007) Does urban sprawl drive changes in the water balance and policy? The case of Leipzig (Germany). 19870–2003. Landscape and Urban Planning 80:1–13

Hellegers P, Soppe R, Perry C, Bastiaanssen W (2010) Remote sensing and economic indicators for supporting water resources management decisions. Water Resour Manag 24(11):2419–2436. doi:10.1007/s11269-009-9559-2

ICWE (1992) (International Conference on Water and Environment). Dublin

Johnson RA, Wichern DW (2002) Applied multivariate statistical analysis. Prentice Hall

Kahn ME (2000) The Environmental Impact of Suburbanization. J Policy Anal Manage 19:569–586

Kanakoudis VK (2002) Urban water use conservation measure. Journal or Water Supply Research and Technology: AQUA 51(3):153–159

Lavière I, Lafrance G (1999) Modelling the electricity consumption of cities: effect of urban density. Energy Econ 21:53–66

Lebart L, Morineau A, Piron M (2006) In Dunod (ed)Statistique exploratoire multidimensionnelle, 4th edn., pp 1–10, 67–142, 2148–184

Liu J, Daily GC, Ehrlich PC, Luck GW (2003) Effects of households dynamics on resource consumption and biodiversity. Nature 421:530–533

Mazzanti Mand Montini A (2006) The determinants of residential water demand: empirical evidence for a panel of Italian municipalities. Appl Econ Lett 13:107–111

Molino B, Rasulo G, Tagliatela L (1996) Forecast model of water consumption for Naples. Water Resour Manag 10(4):321–332. doi:10.1007/BF00508899

Murdock SH, Albrecht DE, Hamm RR, Backman K (1991) Role of sociodemographic characteristics in projections of water use. J Water Resour Plann Manag 117:235–251

Nauges C, Thomas A (2003) Long-run study of residential water consumption. Environmental & Resource Economics, European Association of Environmental and Resource Economists 26(1):25–43

Opaluch JJ (1982) Urban residential demand for water in the United Status: Further discussion. Land Econ 58:225–227

Peña D (2005). Procesos de media móvil y ARMA. In: Alianza (ed). Análisis de series temporales, 1st edn., Madrid, pp 142–163. ISBN 8420691283.

Peña D (2010). El modelo general de regresión. In: Alianza (ed). Regresión y diseño de experimentos, 2nd edn., Madrid, pp 123–149. ISBN 9788420693897.

Postel S (1992). The last oasis. Facing water scarcity. London: W W Norton & Co Inc

Renwick ME, Green (2000) Do residential water demand side management policies measure up? An analysis of eight California water agencies. J Environ Econ Manag 40:37–55

Renzetti S (2002) The economics of water demand. Kluwer Academic Publishers, Boston

Smith A, Ali M (2006) Understanding the impact of cultural and religious water use. Water Environ J 20:203–209

Stephenson D (1999). Demand management theory. Water S.A, numb. 25, 115–122.

United States Environmental Protection Agency (USEPA) (2005) Water and energy savings from high efficiency fixtures and appliances in single family homes. USEPA, Washington

Zhang H, Brown D (2005) Understanding urban residential water use in Beijing and Tianjin, China. Habitat International 29:469–491