Machine Learning for Modeling Water Demand

Maria C. Villarin, Ph.D.[1]; and Victor F. Rodriguez-Galiano, Ph.D.[2]

Abstract: This work shows the application of machine learning (ML) methods to the modeling of water demand for the first time. Classification and regression trees (CART) and random forest (RF), a multivariate, spatially nonstationary and nonlinear ML approach, were used to build a predictive model of water demand in the city of Seville, Spain, at the census tract level. Regression trees (RT) allowed estimation of water demand with an error of 22 L/day/inhabitant and determination of the main driving variables. RF allowed estimation of water demand with error values ranging from 18.89 to 26.91 L/day/inhabitant. The RF method provided better predictions; however, the RT model facilitated better understanding of water demand. This research shows an alternative to the hitherto applied cluster and linear regression approaches for modeling water demand and paves the way for a new set of further scientific investigations based on ML methods. **DOI:** 10.1061/(ASCE)WR.1943-5452.0001067.

Author keywords: Water demand; Census tract; Machine learning; Geography; Urban.

Introduction

Water is a finite resource (Odlare 2014), and water scarcity as a result of overconsumption may affect both the environment and economics (EEA 2015). About 97% of water is in oceans and only 1% of the total amount is available for human consumption (Corcoran et al. 2010), of which 0.7% is used for farming and 0.3% for urban and industrial use (EC 2003). Economic development and population growth have increased the pressure on hydrological resources in the last several decades (Alcamo et al. 2007; Oki et al. 2003; Oki and Kanae 2006; Shiklomanov 2000; Vörösmarty et al. 2000). Mounting water demand has been largely satisfied through the execution of large civil engineering projects for the canalization of watercourses or the creation of dams to store rainwater. These projects prioritized technical and economic needs against environmental criteria (De Nicolás et al. 2014). However, a changed understanding of the relationships between society and nature and of the management of natural resources has also affected water resources, bringing about a change in the traditional hydraulic paradigm (Giansante et al. 2002; Lopez-Gunn 2009). The new management model, called Integrated Water Resources Management (IWRM), incorporates new policy frameworks for administration and decision making. Planners and water managers have forecast water demand and estimated water savings for many conservation programs and measures (Suero et al. 2012). This has resulted in general control and storage water principles ensuring quality water supply to the population and, on the other hand, the environmental sustainability of water systems (EC 2000). With regard to the management of water demand for domestic use, a deep knowledge of the users' behavior relating to water consumption is essential for politicians and public water services managers (Romano et al. 2014). The first initiatives aimed at the study of urban domestic water demand from the perspective of economy, especially in the US, emerge from the 1960s (Conley 1967; Gottlieb 1963; Hanke and Flack 1968; Howe and Linaweaver 1967; Larson and Hudson 1951). These studies analyzed water demand elasticity on the basis of its price. Subsequently, a growing number of studies were developed with more diverse approaches, in which other variables related to weather features (temperature and precipitation) and sociodemographic features (the number of inhabitants per household) were included (Agthe and Billings 1980, 2002; Arbués et al. 2004; Campbell et al. 2004; Dalhuisen et al. 2002; Martinez-Espiñeira 2002; Martínez-Espiñeira and Nauges 2004). Most recently, the range of sociodemographic variables used in this type of study was extended (Villarín 2019). Among the selected variables are those linked to the population age distribution (average age of the population, percentages of both children and teenagers, and even percentage of the elderly, also called the aging ratio) (Fielding et al. 2012; Shandas and Parandvash 2010), building type features (compact or urban sprawl) (March and Saurí 2010), and ethnic features (Inman and Jeffrey 2006; Murdock et al. 1991; Poyer et al. 1997). Hence, in addition to economic factors, territorial, climatic, and technological factors are increasingly present and have become essential factors for planning. In many of these studies the domestic use of water is analyzed at the level of microcomponent per dwelling. Most recently, besides the detailed analysis of water consumption per dwelling, new information supports have been incorporated into the census tract (Ouyang et al. 2014), which allow a more detailed analysis of a given study area. This has resulted in a dramatic increase in information volume and the complexity of water consumption modeling. From this new multiproxy approach, the need has arisen to apply a new generation of computational tools able to extract as much information as possible from increasingly more exhaustive and complex databases.

From a methodological perspective, the modeling of domestic water demand has been carried out through the application of multivariate regression methods such as logistic analysis, linear regression analysis, and hierarchical segmentation analysis (CHAID method), in some cases using factor and cluster analysis (Campbell et al. 1999; Domene and Saurí 2006; Loh 2003; Mayer et al. 1999). On the other hand, in research areas outside the water framework, new algorithms called machine learning (ML) have grown in importance, given the lower restriction in their application and their greater robustness. There are two methods of data modeling: one is based on data stochastic modeling, which has been profusely

[1] Dept. of Human Geography, Univ. of Seville, Seville 41004, Spain (corresponding author). Email: mvillarin@us.es

[2] Physical Geography and Regional Geographic Analysis, Univ. of Seville, Seville 41004, Spain. Email: vrgaliano@us.es

Note. This manuscript was submitted on February 12, 2018; approved on October 22, 2018; published online on March 12, 2019. Discussion period open until August 12, 2019; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Water Resources Planning and Management***, © ASCE, ISSN 0733-9496.

applied in the modeling of water demand; the other is ML, which uses algorithms to generate mechanistic models from learning algorithms (Breiman 2001). The latter comprises inductive knowledge methods with the common denominator of learning patterns from data (data-driven methods). ML has been applied with promising results in other disciplines related to environmental sciences, such as remote sensing (Rodriguez-Galiano et al. 2012a, b), water pollution (Dixon 2009; Rodriguez-Galiano et al. 2014a), and ecology (Archibald et al. 2009; Darling et al. 2012), but the potential for modeling water consumption is still to be explored.

ML is still a relatively new area of science under active development. In the last few decades a great number of methods have emerged. Among the most used are decision trees (Breiman et al. 1984; Leibovici et al. 2011; Qi and Zhu 2011), artificial neural networks (Baykan and Yilmaz 2010; Bue and Stepinski 2006; Canty 2009; Dubois et al. 2007; Mas and Flores 2008; Pavel et al. 2011), support vector machines (Lima et al. 2013; Mountrakis et al. 2011; Petropoulos et al. 2012; Qader et al. 2016; Yu et al. 2012; Zuo and Carranza 2011), and ensembles (Breiman 1996; Rodriguez-Galiano et al. 2014b, 2016), just to mention a few. These methods have different conceptual bases, although they present a series of shared advantages: (1) their capacity to learn complex patterns, taking into account nonlinear relations among the explanatory and dependent variables; (2) high generalization capacity, being robust against incomplete or noisy databases; (3) the possibility of incorporating information a priori; and (4) integration of different types of data in the analysis due to the absence of assumptions about data statistical distributions (for example, normality) (Benediktsson and Sveinsson 1997; Rogan et al. 2003). There are nonetheless tradeoffs. Some of the ML methods (for example, artificial neural networks and support vector machines) behave as black boxes. This means that they can be applied to predict the value of a target variable on the basis of data, but the implicit rules or patterns within the model cannot be interpreted (Coimbra et al. 2014; Tiwari and Adamowski 2015; Yan and Minsker 2011). Within the environmental or social sciences, ML techniques can be very useful for their possibility to be applied on different data types (quantitative variables, ordinal variables, and categorical variables with different distributions). However, it is of key importance to take into account that the final user may not be an ML expert and must be able to interpret results. In this sense, a group of ML algorithms called decision trees provide an alternative to black boxes, by means of the graphical representation of the rules (explanatory variables together with their critical values) that best discriminate subpopulations with different behaviors in the target variable (i.e., water consumption). Recently, ML algorithms based on trees have evolved toward the generation of ensembles, in which prediction is the result of the integration of multiple models (Breiman 1996). Random forest (RF) (Breiman 2001), is probably the greatest expression of this type of algorithm and has been applied to different types of problems successfully (Friedl et al. 1999; Gislason et al. 2006; Rodriguez-Galiano et al. 2015; Sesnie et al. 2008; Steele 2000).

This paper assesses the potential for the application of classification and regression trees (CART) and RF for the modeling of water demand, with the aim of obtaining new information about the potential relationships and interactions between the sociodemographic and urban building characteristics, which might not be identified by more traditional stochastic models. The specific aims were the following:

1. To assess the effectiveness of decision tree algorithms (CART and random forest) for water demand prediction from high-dimensional data;
2. To identify the drivers of water demand in the city of Seville, Spain, and to generate a conceptual graphical model for the different consumption levels at the census tract level; and
3. To develop a feature selection process to eliminate those variables that are not relevant in the modeling of water consumption.

Study Area

The province of Seville, of about 14,036.5 km^2, is located in the region of Andalusia (southern Spain). It is made up of 105 municipalities (IECA 2016). The provincial capital, Seville, is of special importance. The municipality of Seville, hereafter called Seville, has an extension of about 141.3 km^2 (INE 2016). Since 2006, Seville is divided into 11 districts and 522 census sections (AS 2016) (Fig. 1). The River Guadalquivir, the most important river in southern Spain, 640 km in length, crosses the city from north to south (Bhat and Blomquist 2004). The Guadalquivir watershed is 57,527 km^2 in area and runs through 12 provinces belonging to four different regions: Andalusia (90.22%), Castilla-La Mancha (7.13%), Extremadura (2.45%), and Murcia (0.20%) (CHG 2016).

The study area is characterized by a Mediterranean climate, defined by mild temperatures throughout the year and rainfall concentrated during autumn and spring (Appendix S1 in Supplemental Data). The geographical position of the Guadalquivir valley with its low altitude contrasts with its north limit with Sierra Morena-Los Pedroches, and its southeast-northeast limit with the Baetic Mountains. The aperture of the Guadalquivir valley to the Atlantic Ocean makes it possible for west directional component squalls to penetrate, conditioning rainfall distribution with a southeast-northeast direction.

Currently, the company for fresh water supply and wastewater collection of Seville (EMASESA) directly manages fresh water supply and is responsible for public sewerage and water treatment services of the municipality of Seville (EMASESA 2016). The municipality of Seville is supplied by four dams: Aracena (capacity 128.70 hm^3), Zufre (capacity 175.30 hm^3), La Minilla (capacity 57.80 hm^3), and Gergal (capacity 35.00 hm^3) (CHG 2016). In 2016, the Melonares Dam began to operate (capacity 185.6 hm^3). It supplies the municipality of Seville and its metropolitan area with water for human consumption. Considering the data of the volume reached for every dam from October 1999 until September 2016, it is noticeable that the registered volume has fluctuated from 20%–30% in 2000 up to 90%–100% in 2011 (CHG 2016). This shows the strong climate dependence of the volume of water stored and hence the vulnerability of the domestic water supply.

The population in the municipality of Seville for 2009 was 703,206 inhabitants, consisting of 335,097 men and 368,109 women. The average age of both is estimated at 41.4 years (IECA 2016). The percentages of the population over 65 years and under 14 years are 16.35% and 14.87%, respectively. A characteristic spatial pattern for population age distribution can be observed, with the youngest population concentrated in newly built areas and the oldest population in the oldest areas. However, in the central part of the area recuperation and rehabilitation has led to replacing old rental houses with houses that are accessible only to an influx of more affluent inhabitants, a process known as gentrification. Additionally, the number of registered foreign citizens is 71,993 (IECA 2016), mainly from the countries of origin Romania, Morocco, Bolivia, Colombia, and Ecuador. Nationalities from similar geographical environments tend to concentrate in the same areas of the city, which creates similar cultural behaviors regarding water consumption (OECD 1999).

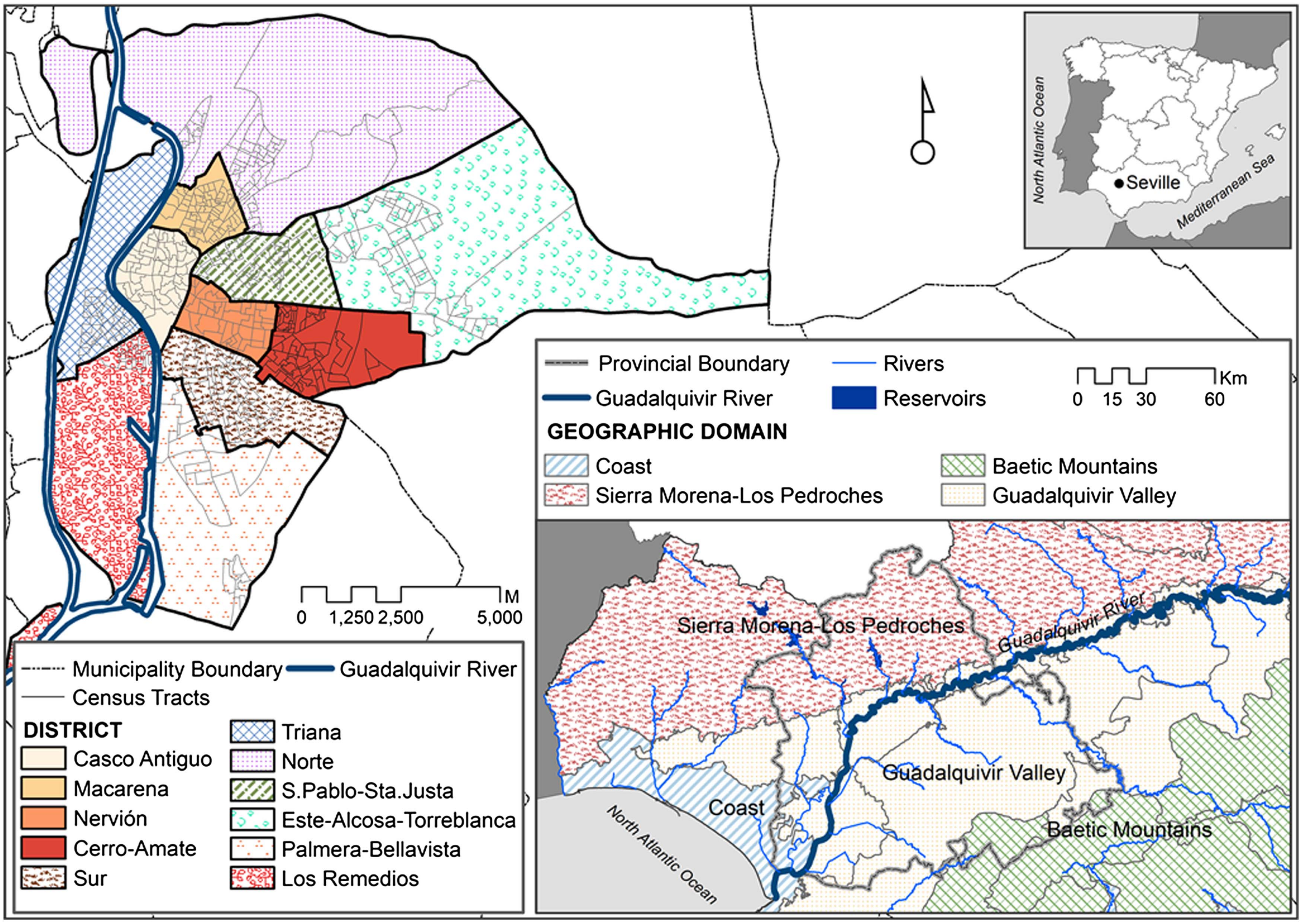


Fig. 1. Study area.

Data

The selected variables for modeling water demand were acquired from several institutions. The mapping base was obtained from 2008 census sections, from the geographical information collection of Andalusian Spatial Data for Intermediate Scales (DEA100), produced and issued by the current Statistics and Mapping Institute of Andalusia (IECA). The original variables, from which the explanatory variables were calculated (Table S1 in Supplemental Data), were represented at various geographic regions or statistical subdivisions defined for the census of Seville city (census tracts). The domestic water consumption (m^3), provided by EMASESA, was computed as the monthly consumption of every client according to the 2009 census tract. The following sociodemographic variables were obtained from the IECA: total population (number of inhabitants), population per age range (number of inhabitants), foreign population (number of inhabitants) and population average age (number of years). The total area (ha) of each census tract was calculated from its polygonal extension, obtained from the DEA100 mapping collection. The surface of the built land (ha), the cadastral surface (m^2), the building height per building (number of floors above ground), the dwellings per building (number of apartments) and the cadastral value per building (€) were obtained from the General Directorate of Cadastre (GDC). Therefore, the total number of final variables introduced in the model amounted to 16, including the domestic water consumption variable. Table 1 provides the basic statistics on these variables, along with their measurement units.

Water consumption was collected on a census tract level after identifying the residential buildings that corresponded to each of its water supply connections. It was therefore possible to relate the monthly invoiced consumption per water supply connection, add the consumption for all water supply connections corresponding to their census tract, and with knowledge of the total population per census tract, calculate the domestic water consumption per capita.

Sociodemographic variables referring to the total population, the population by age range, and the population of foreigners were obtained in absolute values for each census tract. The total population variable was used for subsequent calculation of the population by age range and by number of foreigners, as this allowed conversion of absolute values to relative values in percentages. In the case of the population by age range variable, the principal age ranges were initially established according to the Sauvy (1966) classification as children (0–14 years old), adults (15–64 years old), and the elderly (greater than 65 years old). The adults group (15–64 years old) was further subdivided to create a new group referred to as young adults between 15 and 34, which allowed analysis of their significance in terms of domestic water consumption. In addition to the previously mentioned sociodemographic variables, the youth index and aging index were calculated, given their importance in the processes of population evolution and heir being the main groups of greatest focus in terms of social benefits, as well as the groups that are most sensitive to water consumption (Vinuesa Angulo and Zamora López 1997). The youth and aging indices were calculated as the

Table 1. Basic statistics of the explanatory variables

Variable	Arithmetic average	Standard deviation	Coefficient of variation	Minimum	Maximum
Domestic water consumption (DC) (L/day/inhabitant)	125.38	28.14	0.22 (22%)	14.00	241.73
Population under 15 ($P < 15$) (%)	14.07	4.71	0.33 (33%)	6.94	33.33
Population between 15 and 34 (P1534) (%)	27.27	4.47	0.16 (16%)	16.17	45.59
Population between 35 and 64 (P3564) (%)	40.79	3.91	0.09 (9%)	28.15	50.78
Population over 65 ($P > 65$) (%)	17.91	7.34	0.40 (40%)	1.68	38.85
Youth index (YI) (%)	41.09	4.47	0.10 (10%)	26.10	52.30
Aging index (AI) (%)	5.11	4.34	0.84 (84%)	0.07	38.83
Foreign citizens (FRG) (%)	127.81	187.09	1.46 (146%)	18.47	1,988.89
Average age of population (AAP) (years)	148.23	85.10	0.57 (57%)	5.03	541.30
Average cadastral value (ACV) (€)	4.94	2.10	0.42 (42%)	1.21	11.55
Average built surface area (ABSA) (m^2)	37,453.26	22,708.74	0.60 (60%)	5,150.26	160,078.46
Weighted average height (WAH) (number of floors)	107.52	35.78	0.33 (33%)	42.60	418.59
Average gross density (AGD) ($\text{inhabitants}/m^2$)	256.85	154.23	0.60 (60%)	1.07	974.14
Average net density (AND) [$\text{inhabitants}/m^2$ (constructed)]	65.43	68.59	1.04 (104%)	1.16	653.84
Household size (HS) (inhabitants/household)	2.43	1.28	0.52 (52%)	0.94	28.42
Residential density (RD) [$(\text{inhabitants}/\text{household}) \times 100\ m^2$]	2.44	1.02	0.41 (41%)	0.69	14.90

percentage of the population under the age of 15 relative to the population over the age of 65, and vice versa. The selection of the foreign population variable was based on the specific relationship that can be established between water consumption and the origin of the resident population established in some studies (Inman and Jeffrey 2006; Murdock et al. 1991; Poyer et al. 1997); and the Organisation for Economic Co-operation and Development (OECD). According to the OECD, the decrease in average water consumption is greater for East Asian, Latin American, African, and Indian citizens—up to one-third less than the consumption in Western countries (OECD 1999). Even though this relationship cannot be analyzed in depth on the census tract level, the percentage of foreigners could be considered. The average age of population variable was obtained directly on a census tract level from the IECA.

Dwelling-related variables were obtained from the General Directorate of the Cadastre based on information referring to the cadastral plots of real-estate stock in Seville for 2009. Each cadastral plot classified for residential use was selected per census tract. Thus, the following information was obtained from each census tract and cadastral plot: total area per census tract, residential surface area per census tract, cadastral area per dwelling, building height of each property within each census tract, and the number of dwellings per residential property and their respective cadastral values.

On the basis of the data obtained, the weighted average height (WAH) variable per residential property was calculated

$$\text{WAH} = \frac{\sum_{i=1}^{n} X_i W_i}{\sum_{i=1}^{n} W_i} \tag{1}$$

where X = number of residential buildings; and W = height in floors per residential building.

The increased importance of the number of residential properties was emphasized through the calculation of the weighted average height. Thus, an average value was obtained in the case of mixed census tracts containing a mix of single-family and multifamily dwellings, and a result skewed in favor of taller buildings according to the number of floors was avoided.

The average cadastral value (ACV) variable was calculated on the basis of the average cadastral value per dwelling in each census tract

$$\text{ACV} = \frac{\sum_{i=1}^{n} V_i}{N} \tag{2}$$

where V = cadastral building value; and N = total number of buildings per census tract.

The average built surface area (ABSA) variable per dwelling was obtained on the basis of the cadastral surface area variable per census tract and the number of dwellings

$$\text{ABSA} = \frac{\sum_{i=1}^{n} S_i}{N} \tag{3}$$

where S = residential cadastral surface area; and N = total number of buildings per census tract.

The average gross density (AGD) was obtained as a result of the relationship between the total surface area and total population per census tract

$$\text{AGD} = \frac{p^t}{A} \tag{4}$$

where P^t = total population; and A = surface area by census tract (ha).

The average net density (AND) variable was calculated using the quotient of the residential build surface area and the total number of inhabitants per census tract

$$\text{AND} = \frac{p^t}{\sum_{i=1}^{n} S_i} \tag{5}$$

where P^t = number of inhabitants; and S = residential built surface area by census tract.

Average household size (HS) was obtained using the quotient of the total number of inhabitants and the number of dwellings per census tract

$$\text{HS} = \frac{P^t}{N} \tag{6}$$

where P^t = number of inhabitants; and N = total number of building per census tract.

The residential density (RD) variable was calculated using the quotient of the average number of inhabitants per household and the average cadastral surface area

$$\mathrm{RD} = \frac{P^t}{\sum_{i=1}^{n} S_i} \times 100 \tag{7}$$

where P^t = number of inhabitants; and S = residential cadastral surface area per census tract estimated per every 100 m^2.

Methods

Domestic water demand modeling is a complicated task that can have many different drivers that might not be the same within a given study area. Therefore, assembling a single global model for predicting the water demand of an entire city can be a very confusing and not very realistic goal. Additionally, the drivers or predictive variables for water demand may interact in complicated, nonlinear ways, which can undermine the potential of ordinary statistical techniques. An alternative approach to classical multivariate regression is to subdivide, or partition, the data space into smaller data sets in which the interactions are more manageable. Regression trees (RT) are an alternative to traditional regression (global single predictive models), allowing for multiple nonlinear regressions using recursive partitioning.

The choice to use RT algorithms is usually associated with their simplicity and interpretability, their low computational cost, and the possibility of graphically representing them. Hence, the main benefit of using a hierarchical tree structure to perform regression is that this structure can be viewed as a white box, which in comparison with other machine learning techniques is easier to interpret for understanding the relations between the dependent and independent variable. There exist numerous RT techniques, such as ID3 (Quinlan 1986), C4.5 (Quinlan 1993), CART (Breiman et al. 1984), and others. More recent RT techniques, such us RF (Breiman 2001), have been developed to build ensembles of multiple RTs by repeatedly resampling training data with replacement, and aggregating the trees for a consensus prediction. CART trees and RF have been applied in this study due to their higher performance, interpretability, and availability of implementations in R software.

CART

A decision tree model describes the logical structure of the decisions, uncertainties, and potential outcomes (Khader et al. 2013). CART is composed of a root node, a set of interior nodes, and terminal nodes called leaf nodes. The root node and interior nodes, referred to collectively as nonterminal nodes, are linked into decision stages. The terminal nodes represent the final estimates. Hence, a RT represents a set of restrictions or conditions that are hierarchically organized, which are successively applied from a root to a terminal node or leaf of the tree (Breiman et al. 1984). In order to induce the RT, recursive partitioning and multiple regressions are carried out from the database. From the root node, the data splitting process in each internal node of a rule of the tree is repeated until a previously specified stop condition is reached. Various parameters can be established, such as the minimum number of observations per node, the minimum number of observations in a leaf, and the complexity parameter. Each of the terminal nodes, or leaves, has attached a simple regression model that applies in that node only.

As described by Breiman et al. (1984), the induction of the CART involves first selecting optimal splitting measurement vectors. The process starts by splitting the dependent variable, or the parent node (Dalhuisen et al. 2002), into binary pieces, in which the child nodes are purer than the parent node. Through this process, the CART searches through all candidate splits to find the optimal split, s^*, that maximizes the purity of the resulting tree, as defined by the largest decrease in the impurity

$$\Delta i(s,t) = i(t) - p_{\mathrm{L}} i(t_{\mathrm{L}}) - p_{\mathrm{R}} i(t_{\mathrm{R}}) \tag{8}$$

where s = candidate split at node t. The node t is divided by s into the left child node t_L with a proportion of p_{L}, and the right child node t_{R} with a proportion of p_{R}. Further, $i(t)$ is a measure of impurity before splitting; $i(t_{\mathrm{L}})$ and $i(t_{\mathrm{R}})$ are measures of impurity after splitting; and $\Delta i(s,t)$ measures the decrease in impurity from split s.

There are many approximations for measuring impurity. Some of the most frequently used are gain-ratio (Quinlan 1993), Gini Index (Breiman et al. 1984), and Chi-square (Mingers 1989). CART usually uses the Gini Index as a measure for the best split selection. The Gini index used in this research measures $i(t)$

$$I_{\mathrm{G}}(t_{X(x_i)}) = 1 - \sum_{j=1}^{m} f(t_{X(x_i)}, j)^2 \tag{9}$$

where $f(t_{X(x_i)}, j)$ = proportion of samples with the value x_i belonging to leaf j at node t. The decision tree splitting criterion is based on choosing the attribute with the lowest Gini impurity index (I_{G}).

Random Forest

RF combines the performance of numerous RT algorithms to predict the value of a target variable (Breiman 2001). That is, when RF receives an input vector, (x), made up of the values of the different explanatory variables and the water demand values, RF builds a number K of regression trees and averages the results. After K such trees $\{T(x)\}_1^K$ have been grown, the RF regression predictor is

$$\hat{f}_{rf}^{K}(x) = \frac{1}{K} \sum_{k=1}^{K} T(x) \tag{10}$$

To avoid correlation between the different trees, RF increases the diversity of the trees by making them grow from different training data subsets created through a procedure called bagging. Bagging is a technique used for training data creation by resampling randomly the original data set with replacement, i.e., with no deletion of the data selected from the input sample for generating the next subset $\{h(x, \Theta_k), k = 1, \ldots, K\}$, where $\{\Theta_k\}$ are independent random vectors with the same distribution. Hence, some data may be used more than once in the training, whereas others might never be used. Thus, this process has greater stability, which makes it more robust when facing slight variations in input data, and at the same time it increases prediction accuracy (Breiman 2001). On the other hand, when the RF makes a tree grow, it uses the best feature/split point within a subset of explanatory variables that has been selected randomly from the overall set of input evidential features.

Additionally, the samples that are not selected for the training of the kth tree in the bagging process are included as part of another subset called out-of-bag (oob). These oob elements can be used by the kth tree to evaluate performance (Peters et al. 2007) and to estimate the importance of each explanatory variable in estimating water demand. In this way RF can compute an unbiased estimate of the generalization error without using an external text data subset (Breiman 2001). To assess the importance of each variable (e.g., total population or cadastral building value), the RF switches one of the input evidential features while keeping the rest constant, and it measures the decrease in accuracy that has taken place by means of the oob error estimation (Breiman 2001).

Induction of RT Models

The sociodemographic and urban buildings variables (explanatory variables) and the water consumption values (target variable) were combined into a set of input feature vectors. At each census tract, values from each sociodemographic and urban buildings variable were combined to form a vector. These vectors formed the input to regression tree and random forest algorithms. Water consumption was used as the target values for the induction of the models. Data processing for the induction of the MLA consisted of three stages: (1) training and parameterization of the algorithms; (2) accuracy assessment; and (3) postprocessing requiring converting the output values to a map. All the RT models were created using R 3.2.3 (R-Project) free software. Within this environment, the e1071 library was used for inducting both RT and RF.

For the training of CART it is necessary to set a series of parameters, such as the dissimilarity measure, depth of the tree, and minimum number of observations per node. The dissimilarity measure or heterogeneity influences the way in which the algorithm performs data splits in each node. The depth of the tree and the minimum number of observations are parameters linked to the structural complexity of trees: the greater are the number of levels and the smaller the number of minimum observations in nodes, the greater is the structural complexity of the model. Hence, it is necessary to set these parameters in order to achieve the highest accuracy in prediction and to avoid the creation of complex tree structures that overfit data and lose generality (Pal and Mather 2003). For this study, CART decision-tree models were used (Breiman et al. 1984). For the induction of trees, the Gini index was used as the dissimilarity measure (Breiman et al. 1984; Quinlan 1993). With the aim of obtaining robust and generalizable models, all possible decision trees were assessed, for depth levels of 2 and 3, with a minimum number of observations per node between 30 and 150.

Unlike most other methods based on machine learning, RF needs only two parameters to be set for generating a prediction model: the number of regression trees and the number of evidential features (m), which are used in each node to make regression trees grow (Rodriguez-Galiano et al. 2012b). Breiman (1996) demonstrated that by increasing the number of trees, the generalization error always converges; hence, overtraining is not a problem. On the other hand, reducing the value of m brings as a result a reduction in the correlation among trees, which increases the model's accuracy. In order to optimize these parameters, a large number of experiments were carried out using different number of splits evidential features. The range of the number of trees was set to 2,000, and the number of splits evidential features to between 1 and 15, at intervals of 1.

To assess the optimal value of the different parameters of every method, the predictions derived from all possible parameter combinations were evaluated using the root mean square error (MSE) on the basis of a 10-fold cross-validation procedure. The best model was the one with the lowest RMSE. Relative error (RE) values for each census tract were computed and mapped on the basis of the 10-fold predictions. A feature selection approach, based on the ability of the RF to assess the relative importance of the predictors, was used to identify the minimum number of features that can better explain water demand. To assess the importance of each feature, the RF switched one of the input features while keeping the rest constant, and it re-evaluated the performance of the model measuring the decrease in node impurity (Breiman 2001). The differences were averaged over all 2,000 trees. In order to reduce the number of drivers, the least important feature was removed iteratively at different steps. Then, a 10-fold cross-validation was applied to obtain a stable estimate of the error of the model built after predictor deletions. Finally, the model with a better trade-off between number of predictors and error was chosen as the basis for interpreting the likely drivers of water demand.

Results

Tree-Based Models for Water Consumption

Regression Tree

The most robust RT water consumption model (RMSE and R^2 equal to 22.06 L/day/inhabitant and 0.46, respectively) was obtained by considering a minimum of 36 census tracts in each terminal node and a cost complexity factor of 0.001. This model can be considered robust, given the complexity of the data. Even though all the sociodemographic and urbanization variables were selected as input for the model (mentioned in the section "Data"), RT used only 5 variables in its construction, the most important of which were, in order, HS, ACV, RD, P < 15, and AAP (Table 2). The variables selected for RT showed a significant linear correlation with water consumption in all cases (Table 3). However, the ranking of variables was different—RD, AAP, ACV, P < 15, AI, and so forth—considering a linear regression model. The HS variable, the most important variable in the RT model (Table 3 and Fig. 2), was the variable with the eighth-highest correlation,

Table 2. Variable importance in the RT model

Variable	Importance
HS	20
RD	14
AAP	13
YI	11
P > 65	10
ACV	8
P1534	6
P < 15	6
ABSA	5
AI	5
AND	1

Note: Importance is the frequency of variables in the nodes of the tree, as either the main or the surrogate division variable.

Table 3. Correlation matrix (Pearson correlation coefficient) between variables and DC

Variable	DC
P < 15	-0.398^a
P1534	-0.207^a
P3564	0.04
P > 65	0.360^a
AAP	0.453^a
FRG	0.093^b
YI	-0.287^a
AI	0.369^a
WAH	0.132^a
ACV	0.420^a
ABSA	0.197^a
AGD	0.02
AND	-0.308^a
HS	-0.329^a
RD	-0.479^a

[a]Statistically significant at 99% level.
[b]Statistically significant at 95% level.

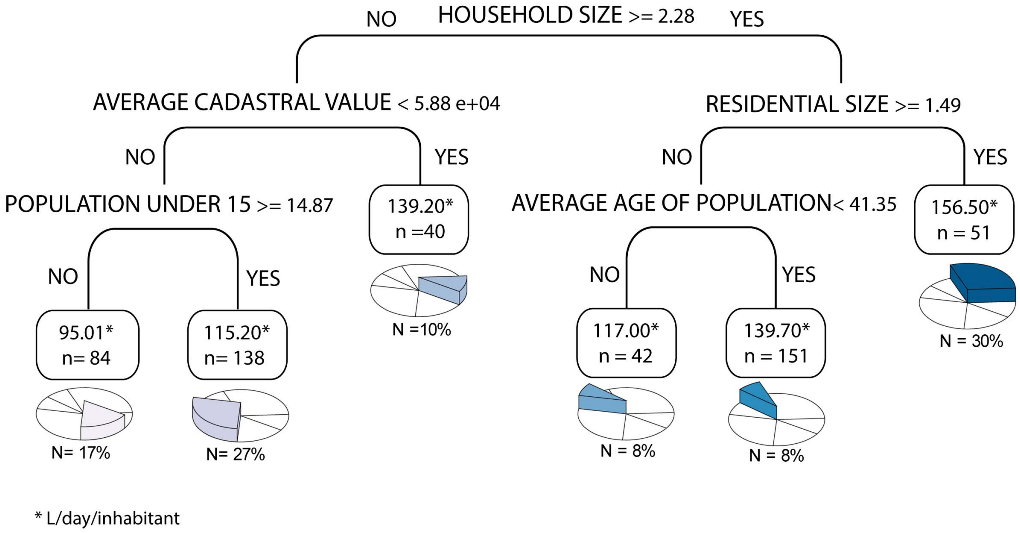


Fig. 2. Decision tree obtained for the census tract. Each explanatory variable is accompanied by the respective threshold value and the percentage of samples included (N). Water consumption values are in L/day/inhabitant.

which could be due to a nonlinear relationship between water consumption and HS.

It should be noted that interpretability of the model played an important role in this research, as it does in many fields of social science. The goal in the application of RT was not to achieve the final results provided by the obtained regression tree (i.e., water consumption estimate) but to better understand the synergies between the remaining variables. The obtained regression tree is shown in Fig. 2, where each socioeconomic variable is accompanied by the respective decision threshold value. The average estimated water consumption is indicated at each terminal node, together with the percentage of the city area to which the estimated values belong. The distribution of samples along the decision tree is not arbitrary but is organized by variables representing different characteristics of the city and its neighborhoods, and the sociourban complexity of it.

The first node or root node was represented by the HS variable. A threshold greater than or equal to 2.28 inhabitants per household was established. When this condition was met, the second node corresponded with the RD variable, for which the threshold was greater than or equal to 1.49 inhabitants per 100 m^2. Therefore, the census tracts with greater domestic consumption (156.50 L/day/inhabitant) were related to a greater number of inhabitants. In this regard, the model followed the patterns detected in other studies that, considering indoor water use only, established a directly proportional relationship with the household size variable in single-family dwellings with 3.35 inhabitants and in multifamily dwellings with 2.19 inhabitants per household (Loh 2003), although in this case no differences were established with regard to the building type. In those census tracts in which RD was less than 1.49 inhabitants per 100 m^2, the model considered a third node represented by AAP for which values lower than 41.35 years of age determined the greatest number of census tracts ($n = 151$) with increased consumption (139.70 L/day/inhabitant). In the case that the foregoing condition was not met, the census tracts presented a lower consumption (117 L/day/inhabitant). The model indicated that consumption increased in census tracts with a population for which the average age was less, a relationship that was also evident in other studies, despite being linked to the fact that children were present.

In census tracts in which HS was less than 2.28 inhabitants per household, the second node was represented by ACV with a lower threshold of €58,800. The cadastral building value was used as a proxy variable for the household income variable, given that the latter information could not be obtained on a census tract level for the year of study. Thus, a greater ACV value signified a greater household income. In those households in which the cadastral value was lower than €55,800, domestic water consumption increased to 139 L/day/inhabitant, although the cadastral value of the buildings was not a variable that was used in other studies on domestic water consumption. The household income variable has traditionally been used since the first econometric studies, such as Larson and Hudson (1951), in which a directly proportional correlation was established between water consumption and the household income variables, which varied between 76 and 190 L/day/inhabitant and $2,000 and $8,000, respectively. This high correlation between income and water consumption was also established in more recent studies (Arbués et al. 2004; Fan et al. 2014; Martinez-Espiñeira 2002; Ojeda de la Cruz et al. 2017). However, in the model that was obtained, the cadastral value was limited to only those households with a HS value of greater than 2.28 inhabitants. It must also be considered that the census tracts obtained were conditioned by their household size rather than by their purchasing power. In contrast, in those census tracts with an ACV of greater than €58,800, a third node was included, represented by $P < 15$ with an established threshold equal to or greater than 14.87%. In those census tracts with large number of minors, the recorded consumption was higher (115 L/day/inhabitant), forming the second-largest group of census tracts ($n = 138$). In census tracts in which the threshold value for $P < 15$ was not reached, domestic water consumption was less (95.01 L/day/inhabitant). In general, less water consumption is expected in an older population (March et al. 2012) and more consumption in a younger population (Campbell et al. 2004). However, other factors have also been established that can limit domestic water consumption, e.g., the existence of water-saving devices or raised awareness regarding water use (Beal et al. 2013; Mamade et al. 2014).

The representation of observed water consumption compared to predicted consumption (Fig. 3) was grouped into four census tract levels in which the observed water consumption varied from 20 L/day/inhabitant up to 240 L/day/inhabitant. The predicted water consumption data for the model varied from 90 L/day/inhabitant up to 160 L/day/inhabitant in a

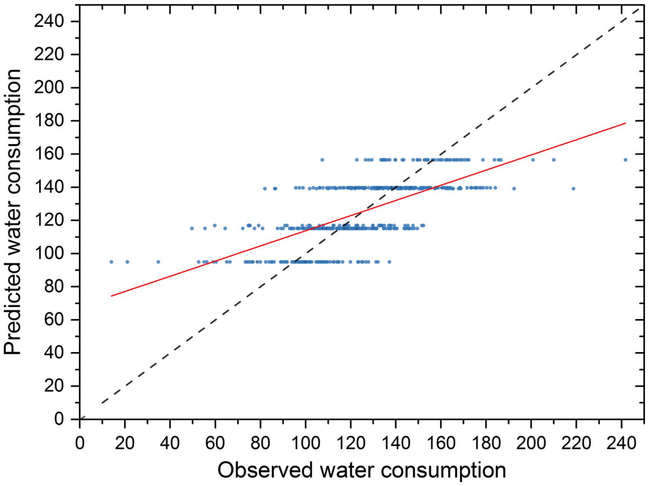


Fig. 3. Observed water consumption compared with RT predictions. The dashed lines represent an exact 1:1 relationship (expected fitting), and the solid lines show a linear regression of these data. The explained variances (percentage R^2) and RMSE values are 46% and 14.01 L/day/inhabitant.

staggered manner. This may be related to the sensitivity or precision of the initial water consumption measurements in some census tracts and to errors in allocating water supply connections to said tracts. EMASESA, the company in charge of supplying water to the city of Seville, collected only the customer contract number associated with each water supply connection on the network, i.e., each water supply point to dwellings. The water supply connections were not georeferenced or associated with census tracts by EMASESA. The association between census tract and cadastral plot was carried out within the framework of this study following the traces of the company's water supply connection network. Therefore, some water supply connections adjacent to various census tracts may have been located in the wrong census tract. Although the total municipal consumption was not affected by this allocation error, it may have affected some of the tracts that share the supply network.

Random Forest Model

The main drivers of water consumption in the city of Seville were identified through the application of a feature selection procedure embedded in the RF method (mentioned in the section "Random Forest"). Fig. 4 shows the RMSE in the prediction of different models after removing the least important variable. RMSE error values ranged from 18.89 L/day/inhabitant to 26.91 L/day/inhabitant. RF produced more robust models than RT, although they were less interpretable, obtaining only the most significant variables and not the rules (gray box). Fig. 5 shows the pseudo-R^2 of the models as well as the relative importance of each explanative variable. RF water consumption models explained a percentage of the variance up to 56%. Regarding the relative importance of the drivers, the same ranking in importance was observed within the different models, which reflected the stability in the RF importance estimation and a high reliability of the results. To interpret the main sociodemographic drivers of the spatial variation in water, a simplified model with a reduced number of variables was selected. The model was composed of 6 variables (pseudo-$R^2 = 0.54$ and RMSE of 18.96 L/day/inhabitant). As in the case of RT, our results suggested that spatial variation in the water consumption of Seville is driven mainly by HS and RD, assigning it much greater importance than to the rest of variables. Therefore, water consumption in Seville is associated mainly with population size. In addition, AAP, ACV, WAH, and $P < 15$ were of significant importance in the model. Fig. 6 shows the relative error box plots for water consumption prediction for all RF models. In all cases, the median relative error value was around 10%, although the maximum error for some census tracts could reach values up to 30% (mentioned in the section "Induction of RT Models"). Additionally, a linear regression between predicted values from RF and observed water consumption produced R^2 values equal to 0.55 and the RMSE value of 13.55 L/day/inhabitant (Fig. 7). However, the lower and higher water consumption values were overestimated and underestimated, respectively.

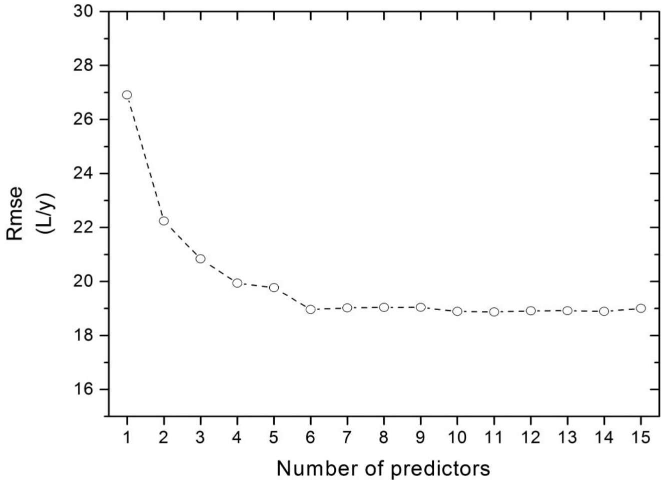


Fig. 4. RMSE of the models fitted as a result of the feature-selection approach.

Spatial Distribution of Modeled Water Consumption

Fig. 8 shows the distribution of predicted and estimated water consumption according to the RT and RF models. The median predicted value for the year of the study was 125.8 L/day/inhabitant. As shown in Fig. 8(a), census tracts with very high values (>160 L/day/inhabitant) corresponded to the Casco Antiguo district (Appendix S2, Fig. S3, and Table S2 in Supplemental Data) in the following neighborhoods: San Lorenzo, Feria, San Gil, Museo, Arenal, and Santa Cruz. This also occurred in a specific number of census tracts in which there was a noticeable presence of single-family dwellings located in the Palmera-Bellavista district (neighborhood of Heliópolis) and the San Pablo–Santa Justa district (neighborhood of Santa Clara); in these census tracts the lowest HS and RD values were recorded. In contrast, census tracts with very low consumption values (<50 L/day/inhabitant) corresponded to those census tracts with a high presence of multifamily dwellings in the Polígono Norte district and to those in which an underestimation of water consumption may occur due to errors in census tract allocation per water supply connection. Census tracts with an average consumption (90–130 L/day/inhabitant) were unevenly distributed and in greater number among the different neighborhoods of the municipality, except in those neighborhoods located in the Casco Antiguo district, in which recorded values were above the average. Census tracts corresponding to lower (50–90 L/day/inhabitant) and higher (130–160 L/day/inhabitant) values than the median value also appeared unevenly. In the case of lower values, these were located in census tracts adjacent to median values, whereas higher values were located in census tracts closer to other higher values. As was observed in the case of higher values in the Casco Antiguo

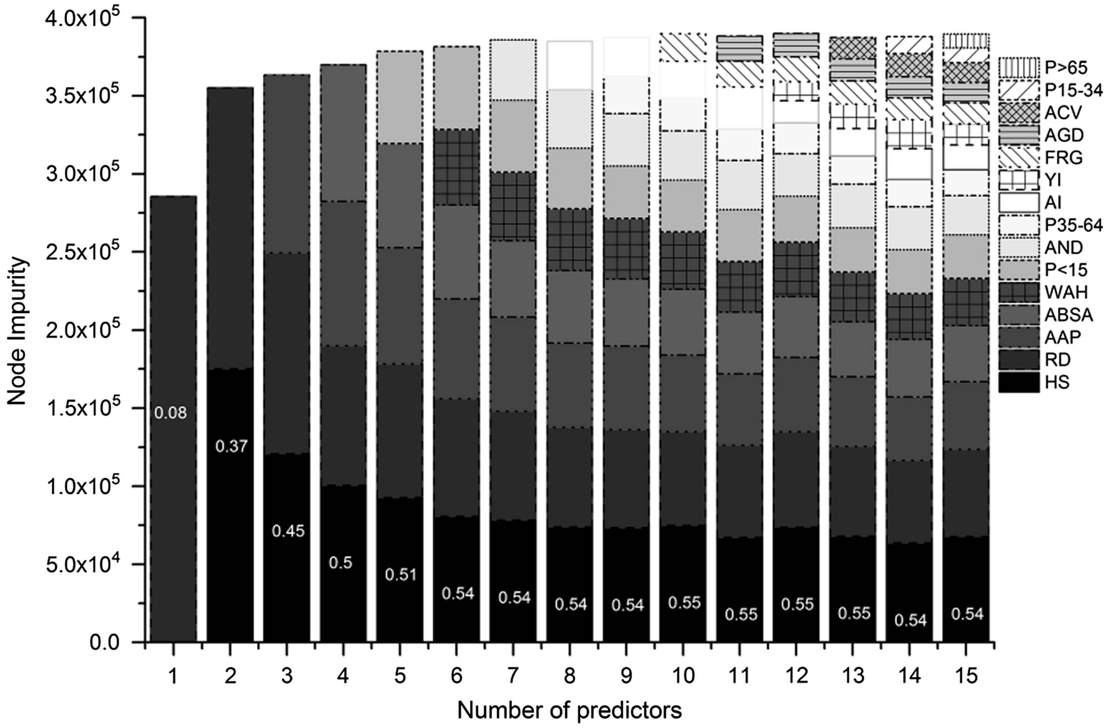


Fig. 5. Relative importance of each independent variable in predicting water consumption in the city of Seville. Different models derived from the feature-selection approach are represented in each column. The number on each column represents the determination coefficients of the model.

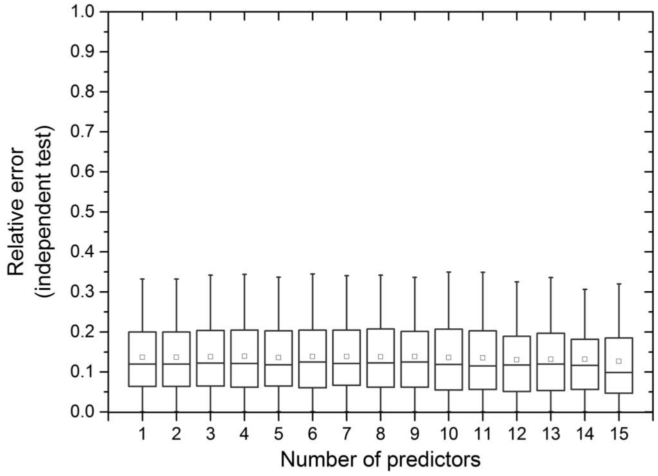


Fig. 6. Relative error of the models fitted as a result of the feature-selection approach: median (interior horizontal line), mean (interior square), 1% and 99% quantiles (edge of boxes), and range (extremes). Relative errors were calculated for a k-fold cross validation test. See Fig. 5 for the explanatory variables in the models, as shown on the x-axis.

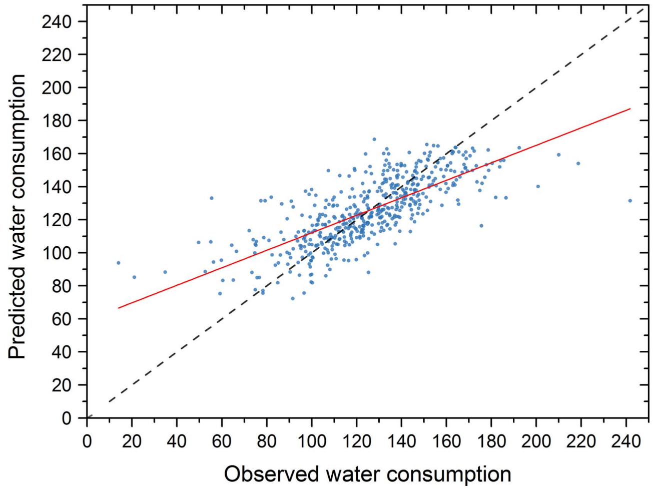


Fig. 7. Observed water consumption compared with the RF predictions calculated using a selection of variables (Fig. 2). The dashed lines represent an exact 1:1 relationship (expected fitting), and the solid lines show a linear regression of these data. The explained variances (percentage R^2) and RMSE values are 55% and 13.55 L/day/inhabitant.

district, these occurred more frequently in the northern area, whereas very high values were located in the southern area.

Figs. 8(b and c) show the domestic water consumption obtained via aggregated machine learning models for intervals that correspond to the average and standard deviations for actual consumption. The RT model [Fig. 8(b)] represents census tracts with only two consumption intervals with average values (90–130 L/day/inhabitant) and high values (130–160 L/day/inhabitant). Consumption predicted by the model placed a high number of census tracts in the average consumption interval, which coincided principally with peripheral areas of the city, whereas neighborhoods with values above the average were located in the Casco Antiguo district as well as in areas of higher economic standing, such as the Los Remedios district (in the neighborhood of the same name), the Nervión district (neighborhood of Buharia), and the San Pablo–Santa Justa district (neighborhoods of Heliópolis and Santa Clara). These last two neighborhoods were also identified in the case of actual domestic consumption.

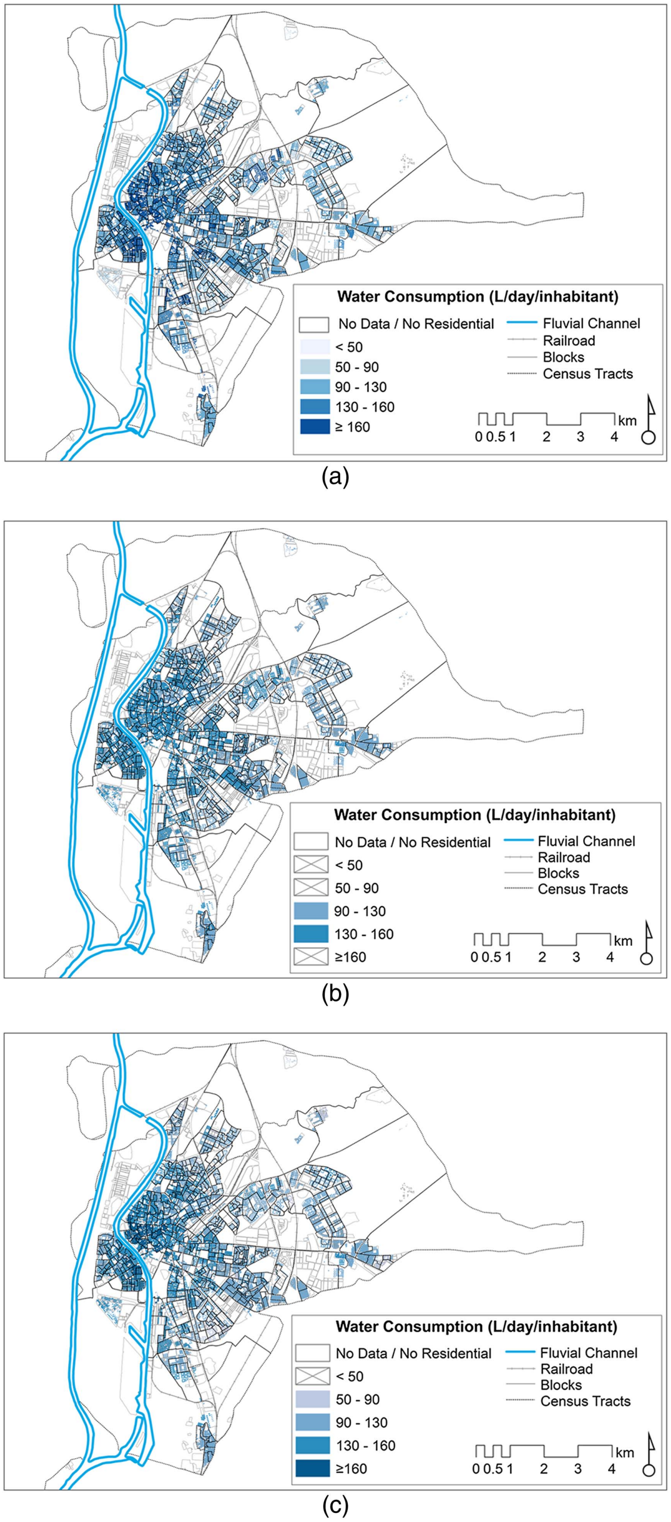


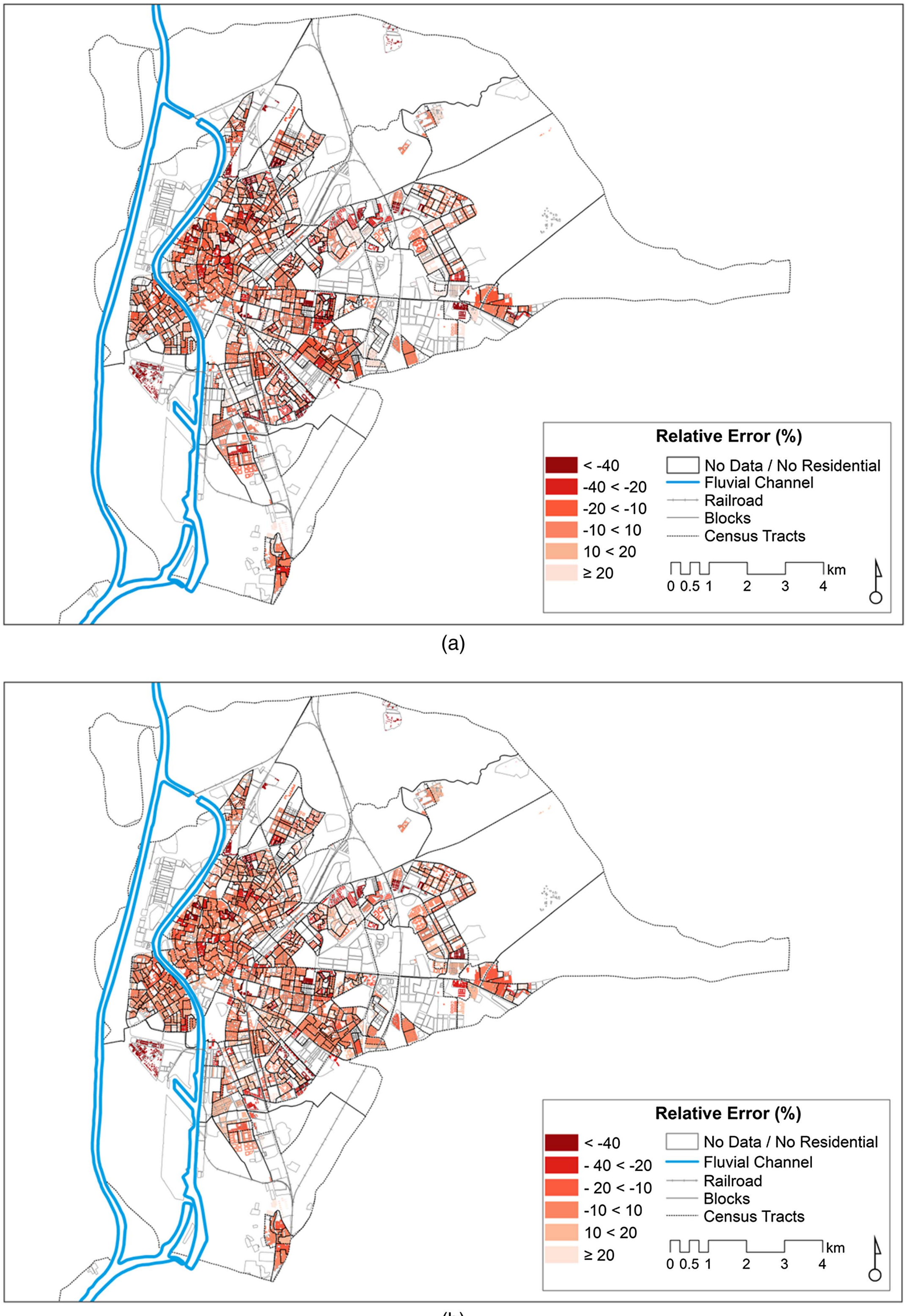
Fig. 8. Domestic water consumption in the city of Seville: (a) real observations; (b) RT model; and (c) RF model.

Fig. 8(c) shows domestic water consumption according to the results obtained according to the RF model. Four intervals were obtained on this occasion. In contrast to actual consumption and the RT model, very low values (<50 L/day/inhabitant) were not associated with any census tract, as can be deduced from the absence of errors in census tract allocation per water supply connection. Census tracts with low consumption (50–90 L/day/inhabitant) were observed in more peripheral areas and corresponded with neighborhoods of lower economic standing, as is the case in the Polígono Norte, Polígono Sur, and Valdezorras neighborhoods. Consumption with average values (90–130 L/day/inhabitant), in line with the RT model, presented a heterogeneous distribution in the municipality, although with an increased trend toward peripheral areas of the city. The neighborhoods of Heliópolis and Santa Clara behaved as expected, as the consumption obtained via the RF model was high (130–160 L/day/inhabitant), precisely as expected for their predominantly single-family composition and in line with the RT model. The census tracts for which very high estimated consumption values were observed (≥160 L/day/inhabitant) were located in the Los Remedios district (neighborhood of Los Remedios) as well as in the southern area of the Casco Antiguo district.

Spatial Distribution of Prediction Models

Fig. 9 shows the relative prediction errors from RT and RF. Relative error values with a negative and positive symbol indicated an underestimation or overestimation of the actual water consumption, respectively, relative to the observed consumption. In the case of the relative error in the RT model [Fig. 9(a) and Table 4], a large number of census tracts were recorded (258, which represents 50.98% of the total) with error values between −10% and 10%. Twenty-nine census tracts underestimated consumption relative to actual consumption by more than 40%. Four census tracts with very low consumption (50 L/day/inhabitant) were obtained in the districts of Macarena (neighborhood of Polígono Sur), Sur (neighborhood of Polígono Sur), and Este-Alcosa-Torreblanca (neighborhood of Palacio de Congresos), possibly due to errors in allocating water supply connections. Other cases of underestimation of actual consumption occurred, such as in the Los Remedios district (neighborhood of Tablada), with a consumption of 81.93 L/day/inhabitant and located in a military base, a fact that could make it difficult to allocate georeferenced water supply connections by EMASESA. Other census tracts with a significant underestimation of actual domestic consumption were located in the Norte district, in the neighborhoods of la Bachillera (64.37 L/day/inhabitant) and El Gordillo (52.56 L/day/inhabitant). In both cases, the errors could be the result of their own historical evolution, given that both increased the number of dwellings in a manner that was not authorized by the Administration and therefore lacking in basic supply and sanitation services. The census tracts that presented an overestimation of actual water consumption relative to expected consumption in the model were located further away from the center of the city. Some of these were observed in the districts of Triana and Los Remedios and registered very high values (above 160 L/day/inhabitant), in which a predominantly multifamily composition with low density should respond to lower actual water consumption. The same specific circumstances were detected in census tracts in the Sur, Macarena, and Este-Alcosa-Torreblanca districts, in which there was abnormally high actual water consumption. However, there were also occasional cases of underestimation of actual water consumption.

In the case of the RF model [Fig. 9(b)], the majority of census tracts were concentrated, in line with the RT model, within the error interval between −10% and 10% (378 census tracts, comprising 74.70% of the total). This signified that, in this last case, the RF model had a better fit with expected consumption in the census tracts. Some census tracts that presented overestimated values coincided with those detected in the RT model, as is the case with the census tracts located in the neighborhoods of Tablada, La Bachillera, and el Gordillo, with the exception of some census tracts located in the Palmera-Bellavista and Este-Alcosa-Torreblanca districts. This supports the error hypothesis in the allocation of water supply connections. It is of note that in this RF model is the



(a)

(b)

Fig. 9. Spatial distribution of relative errors: (a) RT model; and (b) RF model.

underestimation (≥20%) that occurred in the case of the San Pablo–Santa Justa district (neighborhood of Santa Clara), with an actual consumption greater than 160 L/day/inhabitant and that corresponded to the single-family dwelling type and low residential density from which higher actual consumption is expected. The model in this case attributes a consumption of 129.38 L/day/inhabitant, close to the average of actual consumption predicted, 125 L/day/inhabitant. There is only one case in which the census tract registered an estimation error of less than 40%: census section 4027, located in the Los Pájaros neighborhood (Cerro-Amate district), for which actual consumption was above the consumption estimated by the model.

Table 4. Relative error for each RT and RF model on the census tract and neighborhood levels

			Relative error (%)					
Model	Administrative level	Size	<−40	−40 to −20	−20 to −10	−10 to 10	10 to 20	>20
RT	Census tract	N[a]	29	36	66	258	83	34
		%[b]	5.73	7.11	13.04	50.98	16.40	6.71
	Neighborhood	N[c]	1	1	7	95	4	0
		%[d]	0.92	0.92	1.38	87.96	3.70	0
RF	Census tract	N[a]	1	15	55	378	53	4
		%[b]	0.19	2.96	10.86	74.70	10.47	0.79
	Neighborhood	N[c]	1	1	2	104	0	0
		%[d]	0.92	0.92	1.85	96.29	0	0

[a] Number of census tracts.
[b] Percentages of total census tracts.
[c] Number of neighbourhoods.
[d] Percentages of total neighbourhoods.

Conclusions

The study was intended to evaluate the performance of regression tree methods for predictive modeling of water consumption in the city of Seville, Spain, using sociodemographic and urban-building indices as predictors. This research reveals new insights into the drivers of water consumption across a city, while at the same time it establishes a new methodological framework for its predictive modeling. Specifically, the CART and random forest methods, two multivariate, spatially nonstationary, and nonlinear machine learning approaches, were introduced as an alternative to the hitherto applied multiple linear regression approaches, paving the way for further scientific investigation based on machine learning methods.

Tree-based machine learning methods provide many advantages in analyzing domestic water consumption: (1) simplicity in terms of both its application and interpretability of results, (2) the ability to use complex data from different statistical distributions, and (3) recognition of nonlinear relationships between variables. Additionally, the use of feature selection techniques represents an effective tool to better determine the factors that have the highest influence on water consumption and provide guidelines to evaluate their role in influencing high consumption levels. Determining which explanatory variables mainly influence the occurrence of high consumption represents an important step in managing and ensuring both quality water supply to the population and the sustainability of water systems. Our results reveal that multiple sociodemographic and urban-building drivers explain water consumption in the city of Seville. The variables selected for RT showed a significant linear correlation with water consumption in all cases, despite the ranking of variables being different in the linear regression model. RF produced more robust models than RT, but they were less interpretable due to the absence of rules (gray box). Both models placed similar importance on the different variables, which reflects the stability of the estimate and the high reliability of the results. The HS and RD variables had the closest relationship with domestic water consumption recorded in both the RT and the RF models.

Regarding the threshold or cutoff values obtained with RT, it is noteworthy that the census tracts with the greatest domestic consumption (with a maximum of 156.50 L/day/inhabitant) will be related to a HS value greater than 2.28 and a RD value greater than 1.49. In those tracts for which the RD values were below the established threshold, water consumption was determined using the AAP value. The ACV variable was used as a proxy variable for the household income variable. Thus, a greater ACV value signified a greater increase in household incomes. In those dwellings where the ACV was lower than €55,800, domestic water consumption increased (values greater than 139 L/day/inhabitant). Finally, the $P < 15$ variable, for which the threshold was established for values at or greater than 14.87%, indicated that recorded consumption was higher in those census tracts with a larger number of minors (greater than 115 L/day/inhabitant).

The results of the RF model suggest that domestic water consumption will be determined principally by HS and RD, although the AAP, ACV, WAH, and $P < 15$ variables are also of importance in this case. The HS and RD variables showed their highest values in tracts located in the periurban area, identified by new construction and multifamily type dwellings, whereas the lowest values were recorded in census tracts located in older neighborhoods with single-family type dwellings. In terms of the AAP variable, higher values were observed in neighborhoods with less purchasing power and the oldest districts in the municipality. In contrast, the census tracts located in new-construction neighborhoods had lower recorded AAP values. The ACV variable coincided with census tracts with lower HS and RD values. Furthermore, it was confirmed that the lower values recorded for the ACV variable were detected in areas of the city in which the AAP, RD, and HS variables registered elevated values identified with new-construction neighborhoods. The WAH variable showed higher values in census tracts located in neighborhoods that are not included in the center but are consolidated urbanistically. Finally, the $P < 15$ variable showed its highest values in census tracts located in areas of new construction and lower economic value, although not broadly, as values were also recorded that were similar to those observed in the center of the municipality and in areas of greater economic standing.

On one hand, predicted water consumption that was very high (>160 L/day/inhabitant) corresponded with the Casco Antiguo district as well as a specific number of census tracts in which there was a notable presence of single-family dwellings. In contrast, very low consumption (<50 L/day/inhabitant) corresponded to census tracts with a high presence of multifamily dwellings. Census tracts with an average consumption (90–130 L/day/inhabitant) were unevenly distributed and in greater number among the different neighborhoods of the municipality. In the case of the RT model, considering the same intervals as in the case of the predicted consumption, census tracts were observed that coincided with peripheral areas of the city in the case of average values (90–130 L/day/inhabitant), as well as high values (130–160 L/day/inhabitant) belonging to neighborhoods located in the central area with the highest economic value. In the case of the RF model, there were four estimated intervals, with the exception of very low consumption (<50 L/day/inhabitant), that were not present in any census tract, as could be inferred from the lack of error in the allocation of census tracts per water supply connection. Census tracts with low consumption

(50–90 L/day/inhabitant) were observed in the most peripheral areas and corresponded to neighborhoods of lower economic standing. Average (90–130 L/day/inhabitant) and high (130–160 L/day/inhabitant) consumption presented an uneven distribution, as is the case in the RT model. In the case of average consumption, this trended toward peripheral areas, and in the case of higher consumption, this was more present in single-family type dwellings. Census tracts in which very high levels of estimated consumption were observed (≥160 L/day/inhabitant) were located principally in tracts in the southern area of the municipality.

The RF method provides better predictions than RT; however, the interpretability of the RT model facilitates better understanding of existing synergies between predictors and domestic water consumption. The design parameters are simple in the case of RF. The most robust RT water consumption model (RMSE and R^2 equal to 22.06 L/day/inhabitant and 0.46, respectively) was obtained from considering a minimum of 36 census tracts in each terminal node and a cost complexity factor of 0.001, whereas RF water consumption models explained a variance percentage of up to 56%. The model selected comprises six variables (pseudo-$R^2 = 0.54$ and RMSE of 18.96 L/day/inhabitant). In terms of error distribution by census tract, 50.98% of census tracts estimated by RT comprised around 10% of relative error. Furthermore, in the case of the RF model, these tracts represented 74.70% of the total.

Data Availability

The following data, models, or code generated or used during the study are available from the corresponding author by request: demographic, socioeconomic, and building variables and scripts for the application of random forest and CART models.

Acknowledgments

The second author is a Juan de la Cierva grant holder (reference FPDI-2013-17183). The authors are grateful for the financial support given by the Spanish MINECO (project BIA2013-43462-P). We acknowledge the data providers for the water consumption measurements and the sociodemographic and urban building information, EMASESA and IECA, respectively.

Supplemental Data

Appendixes S1 and S2, Figs. S1–S3, and Tables S1 and S2 are available online in the ASCE Library (www.ascelibrary.org).

References

Agthe, D. E., and R. B. Billings. 1980. "Dynamic models of residential water demand." *Water Resour. Res.* 16 (3): 476–480. https://doi.org/10.1029/WR016i003p00476.

Agthe, D. E., and R. B. Billings. 2002. "Water price influence on apartment complex water use." *J. Water Resour. Plann. Manage.* 128 (5): 366–369. https://doi.org/10.1061/(ASCE)0733-9496(2002)128:5(366).

Alcamo, J., M. Flörke, and M. Märker. 2007. "Future long-term changes in global water resources driven by socio-economic and climatic changes." *Hydrol. Sci. J.* 52 (2): 247–275. https://doi.org/10.1623/hysj.52.2.247.

Arbués, F., R. Barberán, and I. Villanúa. 2004. "Price impact on urban residential water demand: A dynamic panel data approach." *Water Resour. Res.* 40 (11): 1–9. https://doi.org/10.1029/2004WR003092.

Archibald, S., D. P. Roy, B. W. van Wilgen, and R. J. Scholes. 2009. "What limits fire? An examination of drivers of burnt area in southern Africa." *Global Change Biol.* 15 (3): 613–630. https://doi.org/10.1111/j.1365-2486.2008.01754.x.

AS (Ayuntamiento de Sevilla). 2016. *Distritos de Sevilla*. Sevilla, Spain: AS.

Baykan, N. A., and N. Yilmaz. 2010. "Mineral identification using color spaces and artificial neural networks." *Comput. Geosci.* 36 (1): 91–97. https://doi.org/10.1016/j.cageo.2009.04.009.

Beal, C. D., R. A. Stewart, and K. Fielding. 2013. "A novel mixed method smart metering approach to reconciling differences between perceived and actual residential end use water consumption." *J. Cleaner Prod.* 60: 116–128. https://doi.org/10.1016/j.jclepro.2011.09.007.

Benediktsson, J. A., and J. R. Sveinsson. 1997. "Feature extraction for multisource data classification with artificial neural networks." *Int. J. Remote Sens.* 18 (4): 727–740. https://doi.org/10.1080/014311697218728.

Bhat, A., and W. Blomquist. 2004. "Policy, politics, and water management in the Guadalquivir River basin, Spain." *Water Resour. Res* 40 (8): W08S071–W08S0711. https://doi.org/10.1029/2003WR002726.

Breiman, L. 1996. "Bagging predictors." *Mach. Learn.* 24 (2): 123–140. https://doi.org/10.1007/BF00058655.

Breiman, L. 2001. "Random forests." *Mach. Learn.* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen. 1984. *Classification and regression trees*. Belmont, CA: Chapman and Hall/CRC.

Bue, B. D., and T. F. Stepinski. 2006. "Automated classification of landforms on Mars." *Comput. Geosci.* 32 (5): 604–614. https://doi.org/10.1016/j.cageo.2005.09.004.

Campbell, H. E., R. M. Johnson, and E. H. Larson. 2004. "Prices, devices, people, or rules: The relative effectiveness of policy instruments in water conservation1." *Rev. Policy Res.* 21 (5): 637–662. https://doi.org/10.1111/j.1541-1338.2004.00099.x.

Campbell, H. E., E. H. Larson, R. M. Johnson, and M. J. Watts. 1999. *Some best bets in residential water conservation: Results of a multivariate regression analysis*, City of Phoenix, 1990–1996. Final Rep. Phoenix, AZ: Morrison Institute for Public Policy.

Canty, M. J. 2009. "Boosting a fast neural network for supervised land cover classification." *Comput. Geosci.* 35 (6): 1280–1295. https://doi.org/10.1016/j.cageo.2008.07.004.

CHG (Confederación Hidrográfica del Guadalquivir). 2016. *S.A.I.H del Guadalquivir.* Andalucía, Spain: CHG.

Coimbra, R., V. Rodriguez-Galiano, F. Olóriz, and M. Chica-Olmo. 2014. "Regression trees for modeling geochemical data—An application to Late Jurassic carbonates (Ammonitico Rosso)." *Comput. Geosci.* 73: 198–207. https://doi.org/10.1016/j.cageo.2014.09.007.

Conley, B. C. 1967. "Price elasticity of the demand for water in Southern California." *Ann. Reg. Sci.* 1 (1): 180–189. https://doi.org/10.1007/BF01290019.

Corcoran, E., C. Nellemann, E. Baker, R. Bos, D. Osborn, H. Savelli, eds. 2010. *Sick water? The central role of waste- water management in sustainable development. A rapid response assessment*. Arendal, Norway: United Nations Environment Programme, UN-HABITAT, GRID-Arendal.

Dalhuisen, J. M., H. L. F. De Groot, C. A. Rodenburg, and P. Nijkamp. 2002. "Economic aspects of sustainable water use: Evidence from a horizontal comparison of European cities." *Int. J. Water* 2 (1): 75–94. https://doi.org/10.1504/IJW.2002.002080.

Darling, E. S., L. Alvarez-Filip, T. A. Oliver, T. R. McClanahan, and I. M. Côté. 2012. "Evaluating life-history strategies of reef corals from species traits." *Ecol. Lett.* 15 (12): 1378–1386. https://doi.org/10.1111/j.1461-0248.2012.01861.x.

De Nicolás, V. L., F. Laguna-Peñuelas, and P. Vidueira. 2014. "An energy optimization criterion for branched water networks." *Tecnologia y Ciencias del Agua* 5 (6): 41–51.

Dixon, B. 2009. "A case study using support vector machines, neural networks and logistic regression in a GIS to identify wells contaminated with nitrate-N." *Hydrogeol. J.* 17 (6): 1507–1520. https://doi.org/10.1007/s10040-009-0451-1.

Domene, E., and D. Saurí. 2006. "Urbanisation and water consumption: Influencing factors in the metropolitan region of Barcelona." *Urban Stud.* 43 (9): 1605–1623. https://doi.org/10.1080/00420980600749969.

Dubois, M. K., G. C. Bohling, and S. Chakrabarti. 2007. “Comparison of four approaches to a rock facies classification problem.” *Comput. Geosci.* 33 (5): 599–617. https://doi.org/10.1016/j.cageo.2006.08.011.

EC (European Communities). 2000. *Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy.* Brussels, Belgium: ATBD X.X, European Communities.

EC (European Communities). 2003. *Water for life. Office for Official Publications of the European Communities*. Luxembourg: European Commission.

EEA (European Environment Agency). 2015. *The European Environment—State and outlook 2015 Report*. Copenhagen, Denmark: EEA.

EMASESA (Empresa Metropolitana de Abastecimiento y Saneamiento de Aguas de Sevilla). 2016. *Área gestionada por EMASESA*. Sevilla, Spain: EMASESA.

Fan, L., G. Liu, F. Wang, C. J. Ritsema, and V. Geissen. 2014. “Domestic water consumption under intermittent and continuous modes of water supply.” *Water Resour. Manage.* 28 (3): 853–865. https://doi.org/10.1007/s11269-014-0520-7.

Fielding, K. S., S. Russell, A. Spinks, and A. Mankad. 2012. “Determinants of household water conservation: The role of demographic, infrastructure, behavior, and psychosocial variables.” *Water Resour. Res.* 48 (10): W10510. https://doi.org/10.1029/2012WR012398.

Friedl, M. A., C. E. Brodley, and A. H. Strahler. 1999. “Maximizing land cover classification accuracies produced by decision trees at continental to global scales.” *Geosci. Remote Sens.* 37 (2): 969–977. https://doi.org/10.1109/36.752215.

Giansante, C., M. Aguilar, L. Babiano, A. Garrido, A. Gómez, E. Iglesias, W. Lise, L. Moral, and B. Pedregal. 2002. “Institutional adaptation to changing risk of water scarcity in the Lower Guadalquivir Basin.” *Nat. Resour. J.* 42 (3): 521–564.

Gislason, P. O., J. A. Benediktsson, and J. R. Sveinsson. 2006. “Random forests for land cover classification.” *Pattern Recognit. Lett.* 27 (4): 294–300. https://doi.org/10.1016/j.patrec.2005.08.011.

Gottlieb, M. 1963. “Urban domestic demand for water: A Kansas case study.” *Land Econ.* 39 (2): 204–210. https://doi.org/10.2307/3144756.

Hanke, S. H., and J. E. Flack. 1968. “Effects of metering urban water.” *J. Am. Water Works Assoc.* 60 (12): 1359–1366. https://doi.org/10.1002/j.1551-8833.1968.tb03685.x.

Howe, C. W., and F. P. Linaweaver. 1967. “The impact of price on residential water demand and its relation to system design and price structure.” *Water Resour. Res.* 3 (1): 13–32. https://doi.org/10.1029/WR003i001p00013.

IECA (Instituto de Estadística y Cartografía de Andalucía). 2016. *Sistema de Información Multiterritorial de Andalucía (SIMA)*. Sevilla, Spain: Instituto de Estadística y Cartografía de Andalucía.

INE (Instituto Nacional de Estadística). 2016. *Revisión del Padrón municipal 2009. Datos por municipios*. Madrid, Spain: INE.

Inman, D., and P. Jeffrey. 2006. “A review of residential water conservation tool performance and influences on implementation effectiveness.” *Urban Water J.* 3 (3): 127–143. https://doi.org/10.1080/15730620600961288.

Khader, A. I., D. E. Rosenberg, and M. McKee. 2013. “A decision tree model to estimate the value of information provided by a groundwater quality monitoring network.” *Hydrol. Earth Syst. Sci.* 17 (5): 1797–1807. https://doi.org/10.5194/hess-17-1797-2013.

Larson, B. O., and H. E. Hudson. 1951. “Residential water use and family income.” *J. Am. Water Works Assn.* 43 (8): 603–611.

Leibovici, D. G., L. Bastin, and M. Jackson. 2011. “Higher-order co-occurrences for exploratory point pattern analysis and decision tree clustering on spatial data.” *Comput. Geosci.* 37 (3): 382–389. https://doi.org/10.1016/j.cageo.2010.06.006.

Lima, A. R., A. J. Cannon, and W. W. Hsieh. 2013. “Nonlinear regression in environmental sciences by support vector machines combined with evolutionary strategy.” *Comput. Geosci.* 50: 136–144. https://doi.org/10.1016/j.cageo.2012.06.023.

Loh, M. C. P. 2003. *Domestic water use study in Perth, Western Australia 1998–2001*. Perth, Australia.

Lopez-Gunn, E. 2009. “Agua para todos: A new regionalist hydraulic paradigm in Spain.” *Water Altern.* 2 (3): 370–394.

Mamade, A., D. Loureiro, D. Covas, S. T. Coelho, and C. Amado. 2014. “Spatial and temporal forecasting of water consumption at the DMA level using extensive measurements.” *Procedia Eng.* 70: 1063–1073. https://doi.org/10.1016/j.proeng.2014.02.118.

March, H., J. Perarnau, and D. Saurí. 2012. “Exploring the links between immigration, ageing and domestic water consumption: The case of the metropolitan area of Barcelona.” *Reg. Stud.* 46 (2): 229–244. https://doi.org/10.1080/00343404.2010.487859.

March, H., and D. Saurí. 2010. “The suburbanization of water scarcity in the Barcelona metropolitan region: Sociodemographic and urban changes influencing domestic water consumption.” *Prof. Geogr.* 62 (1): 32–45. https://doi.org/10.1080/00330120903375860.

Martinez-Espiñeira, R. 2002. “Residential water demand in the northwest of Spain.” *Environ. Resour. Econ.* 21 (2): 161–187. https://doi.org/10.1023/A:1014547616408.

Martínez-Espiñeira, R., and C. Nauges. 2004. “Is all domestic water consumption sensitive to price control?” *Appl. Econ.* 36 (15): 1697–1703. https://doi.org/10.1080/0003684042000218570.

Mas, J. F., and J. J. Flores. 2008. “The application of artificial neural networks to the analysis of remotely sensed data.” *Int J. Remote Sens.* 29 (3): 617–663. https://doi.org/10.1080/01431160701352154.

Mayer, P. W., W. B. DeOreo, E. M. Opitz, J. C. Kiefer, W. Y. Davis, B. Dziegielewski, and J. O. Nelson. 1999. *Residential end uses of water*, 352. Denver: American Water Works Association.

Mingers, J. 1989. “An empirical comparison of selection measures for decision-tree induction.” *Mach. Learn.* 3 (4): 319–342. https://doi.org/10.1007/BF00116837.

Mountrakis, G., J. Im, and C. Ogole. 2011. “Support vector machines in remote sensing: A review.” *ISPRS J. Photogramm. Remote Sens.* 66 (3): 247–259. https://doi.org/10.1016/j.isprsjprs.2010.11.001.

Murdock, S. H., D. E. Albrecht, R. R. Hamm, and K. Backman. 1991. “Role of sociodemographic characteristics in projections of water use.” *J. Water Resour. Plann. Manage.* 172 (2): 235–251. https://doi.org/10.1061/(ASCE)0733-9496(1991)117:2(235).

Odlare, M. 2014. “Introductory chapter for water resources.” In *Reference module in earth systems and environmental sciences.* New York: Elsevier.

OECD. 1999. *The price of water.* OECD.

Ojeda de la Cruz, A., C. R. Alvarez-Chavez, M. A. Ramos-Corella, and F. Soto-Hernandez. 2017. “Determinants of domestic water consumption in Hermosillo, Sonora, Mexico.” *J. Cleaner Prod.* 142: 1901–1910. https://doi.org/10.1016/j.jclepro.2016.11.094.

Oki, T., Y. Agata, S. Kanae, T. Saruhashi, and K. Musiake. 2003. “Global water resources assessment under climatic change in 2050 using TRIP.” In *Proc., 280 of IAHS-AISH Publication*, 124–133. Wallingford, Oxfordshire, UK: International Association of Hydrological Sciences.

Oki, T., and S. Kanae. 2006. “Global hydrological cycles and world water resources.” *Science* 313 (5790): 1068–1072. https://doi.org/10.1126/science.1128845.

Ouyang, Y., E. A. Wentz, B. L. Ruddell, and S. L. Harlan. 2014. “A multi-scale analysis of single-family residential water use in the phoenix metropolitan area.” *JAWRA J. Am. Water Resour. Assoc.* 50 (2): 448–467. https://doi.org/10.1111/jawr.12133.

Pal, M., and P. M. Mather. 2003. “An assessment of the effectiveness of decision tree methods for land cover classification.” *Remote Sens. Environ.* 86 (4): 554–565. https://doi.org/10.1016/S0034-4257(03)00132-9.

Pavel, M., J. D. Nelson, and R. Jonathan Fannin. 2011. “An analysis of landslide susceptibility zonation using a subjective geomorphic mapping and existing landslides.” *Comput. Geosci.* 37 (4): 554–566. https://doi.org/10.1016/j.cageo.2010.10.006.

Peters, J., B. De Baets, N. E. C. Verhoest, R. Samson, S. Degroeve, P. De Becker, and W. Huybrechts. 2007. “Random forests as a tool for ecohydrological distribution modelling.” *Ecol. Modell.* 207 (2–4): 304–318. https://doi.org/10.1016/j.ecolmodel.2007.05.011.

Petropoulos, G. P., C. Kalaitzidis, and K. Prasad Vadrevu. 2012. “Support vector machines and object-based classification for obtaining land-use/cover cartography from Hyperion hyperspectral imagery.” *Comput. Geosci.* 41: 99–107. https://doi.org/10.1016/j.cageo.2011.08.019.

Poyer, D. A., L. Henderson, and A. P. S. Teotia. 1997. "Residential energy consumption across different population groups: Comparative analysis for Latino and non-Latino households in USA." *Energy Econ.* 19 (4): 445–463. https://doi.org/10.1016/S0140-9883(97)01024-4.

Qader, S. H., J. Dash, P. M. Atkinson, and V. Rodriguez-Galiano. 2016. "Classification of vegetation type in Iraq using satellite-based phenological parameters." *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 9 (1): 414–424. https://doi.org/10.1109/JSTARS.2015.2508639.

Qi, F., and A. X. Zhu. 2011. "Comparing three methods for modeling the uncertainty in knowledge discovery from area-class soil maps." *Comput. Geosci.* 37 (9): 1425–1436. https://doi.org/10.1016/j.cageo.2010.10.016.

Quinlan, J. R. 1986. "Induction of decision trees." *Mach. Learn.* 1 (1): 81–106. https://doi.org/10.1007/BF00116251.

Quinlan, J. R. 1993. *C4.5 Programs for Machine Learning*. San Francisco: Morgan Kaufmann.

Rodriguez-Galiano, V., M. P. Mendes, M. J. Garcia-Soldado, M. Chica-Olmo, and L. Ribeiro. 2014a. "Predictive modeling of groundwater nitrate pollution using random forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (southern Spain)." *Sci. Total Environ.* 476: 189–206. https://doi.org/10.1016/j.scitotenv.2014.01.001.

Rodriguez-Galiano, V., M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas. 2015. "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines." *Ore Geol. Rev.* 71: 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001.

Rodriguez-Galiano, V. F., M. Chica-Olmo, F. Abarca-Hernandez, P. M. Atkinson, and C. Jeganathan. 2012a. "Random forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture." *Remote Sens. Environ.* 121: 93–107. https://doi.org/10.1016/j.rse.2011.12.003.

Rodriguez-Galiano, V. F., M. Chica-Olmo, and M. Chica-Rivas. 2014b. "Predictive modelling of gold potential with the integration of multisource information based on random forest: A case study on the Rodalquilar area, southern Spain." *Int. J. Geogr. Inf. Sci.* 28 (7): 1336–1354. https://doi.org/10.1080/13658816.2014.885527.

Rodriguez-Galiano, V. F., B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez. 2012b. "An assessment of the effectiveness of a random forest classifier for land-cover classification." *ISPRS J. Photogramm. Remote Sens.* 67 (1): 93–104. https://doi.org/10.1016/j.isprsjprs.2011.11.002.

Rodriguez-Galiano, V. F., M. Sanchez-Castillo, J. Dash, P. M. Atkinson, and J. Ojeda-Zujar. 2016. "Modelling interannual variation in the spring and autumn land surface phenology of the European forest." *Biogeosciences* 13 (11): 3305–3317. https://doi.org/10.5194/bg-13-3305-2016.

Rogan, J., J. Miller, D. Stow, J. Franklin, L. Levien, and C. Fischer. 2003. "Land-cover change monitoring with classification trees using landsat TM and ancillary data." *Photogramm. Eng. Remote Sens.* 69 (7): 793–804. https://doi.org/10.14358/PERS.69.7.793.

Romano, G., N. Salvati, and A. Guerrini. 2014. "Estimating the determinants of residential water demand in Italy." *Forests* 5 (9): 2929–2945. https://doi.org/10.3390/w6102929.

Sauvy, A. 1966. *Théorie générale de la population. II, II*. Paris: Presses Universitaires de France.

Sesnie, S. E., P. E. Gessler, B. Finegan, and S. Thessler. 2008. "Integrating Landsat TM and SRTM-DEM derived variables with decision trees for habitat classification and change detection in complex neotropical environments." *Remote Sens. Environ.* 112 (5): 2145–2159. https://doi.org/10.1016/j.rse.2007.08.025.

Shandas, V., and G. H. Parandvash. 2010. "Integrating urban form and demographics in water-demand management: An empirical case study of Portland, Oregon." *Environ. Plann. B Plann. Des* 37 (1): 112–128. https://doi.org/10.1068/b35036.

Shiklomanov, I. A. 2000. "Appraisal and assessment of world water resources." *Water Int.* 25 (1): 11–32. https://doi.org/10.1080/02508060008686794.

Steele, B. M. 2000. "Combining multiple classifiers: An application using spatial and remotely sensed information for land cover type mapping." *Remote Sens. Environ.* 74 (3): 545–556. https://doi.org/10.1016/S0034-4257(00)00145-0.

Suero, F. J., P. W. Mayer, and D. E. Rosenberg. 2012. "Estimating and verifying United States households' potential to conserve water." *J. Water Resour. Plann. Manage.* 138 (3): 299–306. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000182.

Tiwari, M. K., and J. F. Adamowski. 2015. "Medium-term urban water demand forecasting with limited data using an ensemble wavelet-bootstrap machine-learning approach." *J. Water Resour. Plann. Manage.* 141 (2): 04014053. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000454.

Villarín, M. C. 2019. "Methodology based on fine spatial scale and preliminary clustering to improve multivariate linear regression analysis of domestic water consumption." *Appl. Geogr.* 103: 22–39. https://doi.org/10.1016/j.apgeog.2018.12.005.

Vinuesa Angulo, J., and F. Zamora López. 1997. *Demografía: análisis y proyecciones*. Madrid, Spain: Editorial Síntesis.

Vörösmarty, C. J., P. Green, J. Salisbury, and R. B. Lammers. 2000. "Global water resources: Vulnerability from climate change and population growth." *Science* 289 (5477): 284–288. https://doi.org/10.1126/science.289.5477.284.

Yan, S., and B. Minsker. 2011. "Applying dynamic surrogate models in noisy genetic algorithms to optimize groundwater remediation designs." *J. Water Resour. Plann. Manage.* 137 (3): 284–292. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000106.

Yu, L., A. Porwal, E. J. Holden, and M. C. Dentith. 2012. "Towards automatic lithological classification from remote sensing data using support vector machines." *Comput. Geosci.* 45: 229–239. https://doi.org/10.1016/j.cageo.2011.11.019.

Zuo, R., and E. J. M. Carranza. 2011. "Support vector machine: A tool for mapping mineral prospectivity." *Comput. Geosci.* 37 (12): 1967–1975. https://doi.org/10.1016/j.cageo.2010.09.014.