Case Study

Urban Water Demand Modeling Using Machine Learning Techniques: Case Study of Fortaleza, Brazil

Taís Maria Nunes Carvalho, S.M.ASCE[1]; Francisco de Assis de Souza Filho, D.Sc.[2]; and Victor Costa Porto[3]

Abstract: Despite recent efforts to apply machine learning (ML) for water demand modeling, overcoming the black-box nature of these techniques to extract practical information remains a challenge, especially in developing countries. This study integrated random forest (RF), self-organizing map (SOM), and artificial neural network (ANN) techniques to assess water demand patterns and to develop a predictive model for the city of Fortaleza, Brazil. We performed the analysis at two spatial scales, with different level of information: census tracts (CTs) at the fine scale, and census blocks (CBs) at the coarse scale. At the CB scale, demand was modeled with socioeconomic, demographic, and household characteristics. The RF technique was applied to rank these variables, and the most relevant were used to cluster census blocks with SOMs. RFs and ANNs were used in an iterative approach to define the input variables for the predictive model with minimum redundancy. At the CT scale, demand was modeled using HDI and per capita income. Variables which assess the education level and economic aspects of households demonstrated a direct relationship with water demand. The analysis at the coarse scale provided more insight into the relationship between the variables; however, the predictive model performed better at the fine scale. This study demonstrates how data-driven models can be helpful for water management, especially in environments with strong socioeconomic inequalities, where urban planning decisions should be integrated and inclusive. **DOI: 10.1061/(ASCE)WR.1943-5452.0001310.**

Introduction

The management of water resources systems in rapidly urbanized cities is challenging, especially in regions with high climate variability. Domestic water use is expected to grow significantly over the next two decades in nearly all regions of the world, except in some cities in developed countries (UNESCO 2018; Sauri forthcoming). Freshwater availability will remain constant or decrease (UNESCO 2018), increasing the competition for water and the vulnerability of water supply systems. The risk of water scarcity requires strategies of water conservation or capacity expansion, with the inclusion of alternative water sources. Accurate prediction of water demand is crucial for effective long-term planning. However, water demand is driven by complex, nonlinear interactions between human and ecological systems that are not fully understood (House-Peters and Chang 2011). Previous studies showed that socioeconomic aspects influence domestic water use (Matos et al. 2014; Nawaz et al. 2019), but this relationship is distinct in each region.

Fortaleza, Brazil has a history of multiyear droughts and water supply crisis. The city is supplied by multiple surface water reservoirs, which also are used for irrigation and industrial purposes. Annual precipitation is low and highly variable; hence, water availability is subject to climate conditions. To expand the supply system's capacity and to reduce its climate dependence, local managers plan to install a desalination and wastewater reuse plants. The capacity expansion plan consists of scheduled decisions about when and which source to use in the next 30 years. Research is needed to better understand how the complex interactions between socioeconomic changes and water demand may develop over the coming decades. Currently, managers predict water demand based only on estimated population growth and the average income of the neighborhoods. However, this approach neglects social heterogeneity in the neighborhoods and other aspects that might influence water use (e.g., education and household composition). This study provides a framework for water demand modeling using machine learning techniques and explored the influence of socioeconomic variables on the average daily consumption across Fortaleza.

There is a lack of studies that assess domestic water demand in developing countries, where research is needed to develop social-aware water allocation strategies (UNESCO-WWAP 2019). Domestic water consumption in Brazil was explored in a few previous studies (Brentan et al. 2017; Dias et al. 2018; Sant'Ana and Mazzega 2018; Garcia et al. 2019). However, they were limited to the midwest and southern regions, which have a very different climate and social context from Northeast Brazil.

Outside Brazil, different approaches have been used for water demand modeling, such as regression-type methods, e.g., independent component regression (Haque et al. 2017), multiple linear and evolutionary polynomial regression (Hussien et al. 2016), ordinary least-squares regression (Nawaz et al. 2019), Bayesian linear regression (Rasifaghihi et al. 2020), linear mixed-effects (Romano et al. 2014), autoregressive moving average (Gharabaghi et al. 2019), and agent-based (Xiao et al. 2018) models. Machine learning (ML) techniques have received increasing attention as researchers have come to understand that these algorithms effectively can learn information from water demand data and capture nonlinear relationships between water demand and relevant variables. Lee and Derrible (2020) and Bolorinos et al. (2020) showed that ML models outperform linear methods for prediction of

[1]Ph.D. Student, Dept. of Hydraulic and Environmental Engineering, Universidade Federal do Ceará, Campus do Pici, Bloco 713, CEP 60455-760, Fortaleza, Brazil (corresponding author). ORCID: https://orcid.org/0000-0001-8658-9781. Email: taismarianc@gmail.com

[2]Professor, Dept. of Hydraulic and Environmental Engineering, Universidade Federal do Ceará, Campus do Pici, Bloco 713, CEP 60455-760, Fortaleza, Brazil. Email: assissouzafilho@gmail.com

[3]Ph.D. Student, Dept. of Hydraulic and Environmental Engineering, Universidade Federal do Ceará, Campus do Pici, Bloco 713, CEP 60455-760, Fortaleza, Brazil. Email: victorcporto@gmail.com

Note. This manuscript was submitted on October 7, 2019; approved on August 7, 2020; published online on October 31, 2020. Discussion period open until March 31, 2021; 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.

residential water demand. Duerr et al. (2018) showed that ML can be useful to quantify long-term uncertainty in water demand predictions. Data mining techniques also have been applied to customer segmentation, i.e., to characterize groups of water users, using smart meter data (Cardell-Oliver et al. 2016; Cominola et al. 2018, 2019; Bolorinos et al. 2020).

The most popular machine learning methods in water demand studies are artificial neural networks (ANNs), which long have been used because of their excellent predictive ability (Vijai and Sivakumar 2018). Prior research explored ANN models for predicting 15-min (Guo et al. 2018), weekly (Bata et al. 2020; Adamowski and Karapataki 2010), and monthly water demand (Firat et al. 2008; Altunkaynak and Nigussie 2017); residential water end use (Bennett et al. 2013); and irrigation demand (Pulido-Calvo et al. 2007). Other studies combined ANNs with different methods to improve water demand prediction, such as seasonal autoregressive integrated moving average (Bata et al. 2020) and discrete wavelet transform (Altunkaynak and Nigussie 2017).

Alternative ML techniques used to model water demand are support vector machines (Msiza et al. 2007; Brentan et al. 2017); genetic programming (Liu et al. 2015; Yousefi et al. 2017); and tree-based methods, such as regression trees and random forests (RFs) (Villarin and Rodriguez-Galiano 2019; Bolorinos et al. 2020).

Random forest algorithms stand out in water science and hydrological applications (Tyralis et al. 2019). They have been used mainly for streamflow and water quality modeling (Yajima and Derot 2017; Papacharalampous and Tyralis 2018). A few researchers applied this method to analyze variable importance for water demand prediction (Villarin and Rodriguez-Galiano 2019; Brentan et al. 2017) and short-term forecasting (Vijai and Sivakumar 2018; Chen et al. 2017; Herrera et al. 2010).

ML techniques also are useful for pattern recognition. Self-organizing maps (SOMs)—a type of neural network—have been used in several water resources applications, such as ground water–level forecasting modeling (Haselbeck et al. 2019), water quality assessment (Li et al. 2018), and analysis of land-use changes with satellite data (Qi et al. 2019). SOMs also were used to analyze water consumption patterns (Brentan et al. 2017; Padulano and Giudice 2018).

The modeling approach depends on the data available and the planning horizon. ML methods are useful due to the lack of understanding of the underlying processes driving water demand (Solomatine et al. 2009), but are sensitive to the data set size and the choice of input variables. Lee and Derrible (2020) investigated the role of data availability in water demand modeling; ML models performed better when a larger number of explanatory variables were considered. However, increasing the number of input variables means increasing the number of model parameters, which could reduce the accuracy of the model (Guo et al. 2018). Hence, variable selection is an important step in the modeling process if the data set is extensive.

Long-term prediction usually is related to structural, social and environmental variables, such as lot size, building density, educational level, and family size (Chang et al. 2010; Polebitski and Palmer 2010). Social and structural dynamics might influence changes in water-use behavior, as indicated by Gonzales and Ajami (2017). Understanding these relationships is helpful for tailoring demand-side management strategies and drought-related public measures (Hemati et al. 2016; Lindsay et al. 2017; Quesnel and Ajami 2017). However, this discussion has been limited mostly to the US and Europe.

This study provides further insight into the application and interpretation of machine learning methods for water demand modeling, considering the implications of data availability and spatial level aggregation on model performance. Previous studies focused on evaluating the predictive power of ML models, and so far, there has been little discussion of the individual effect of sociodemographic variables on water demand, especially in developing countries. We addressed this issue with the application the accumulated local effects method (Apley and Zhu 2016) for interpreting black-box models. Domestic water demand was analyzed with cross-sectional data at two spatial levels, the census tract (CT) and the census block (CB). Whereas at the census tract level (fine scale), only 2 variables were available, at the census block level (coarse scale), 18 explanatory variables were used. RFs were used to rank the variables, and SOMs were used to cluster water demand based on the sociodemographic variables. This approach allows the evaluation of possible shifts in water consumption patterns based on socioeconomic scenarios. A predictive model using an ANN was built for both spatial levels. At the census block level, the iterative input selection (IIS) method (Galelli and Castelletti 2013) was used to select the input variables for the predictive model.

Study Area

The city of Fortaleza, capital of Ceará, is in the Northeast region of Brazil. Fortaleza is part of the Metropolitan Region of Fortaleza, which comprises 19 municipalities of Ceará. The territory is divided into 119 neighborhoods, 3,043 census tracts, and 247 census blocks [Fig. 1(b)]. The city is the fifth most populated in Brazil and has the highest demographic density, with over 2.6 million inhabitants distributed across 314.9 km^2. There are 88 men per 100 women; 22.58% of the population is under 14 years old, and 6.58% of the population is over 65 years old. The population is irregularly distributed: the number of inhabitants per neighborhood ranges from 1,000 to 76,000; the most populous neighborhoods are in the south, southeast, and northwest of Fortaleza.

The main river is Cocó, 50 km in length, which crosses the city from north to south (changing to east-southwest) and drains about 60% of the water collected in the Metropolitan Region of Fortaleza into the Atlantic Ocean. The Cocó watershed is the largest in the city (485 km^2) and has 18.7 km^2 of vegetation, including mangroves, dunes, and cerrado (Brazilian savanna). The coastline is 34 km long, and the coastal plain has elevations of less than 200 m.

Fortaleza is characterized by a tropical wet and dry climate, with an average monthly temperature between 24°C and 30.7°C (IPLANFOR 2015). Interannual variability of annual precipitation ranges between 500 and 2,800 mm, and 70% of the total precipitation is concentrated in three months (February–April). The state of Ceará has a long history of multiyear droughts, aggravated by elevated evaporation rates and hydrogeological conditions unfavorable to groundwater storage. All these factors result in low water availability and a vulnerable water supply system.

The local water company, Water and Wastewater Company of Ceará (CAGECE), manages water supply and wastewater collection and treatment for the city. Fortaleza is supplied by eight storage reservoirs, pump stations, and canals that transfer water from the Jaguaribe River basin, through the Jaguaribe-Metropolitano hydrosystem [Fig. 1(a)]. Five of these reservoirs supply Fortaleza, corresponding to a capacity of 871 hm^3, and the other three supply the Jaguaribe basin, with a storage capacity of 10,241 hm^3. Water use in the Jaguaribe region is mainly for irrigation, which accounts for 71% of the system's total demand.

The total water demand of the Jaguaribe-Metropolitano system is estimated at 45.30 m^3/s. The metropolitan basin main uses are domestic, municipal, and industrial. The western part of Fortaleza is the main industrial area of the region, with a water consumption of 1.4 m^3/s.

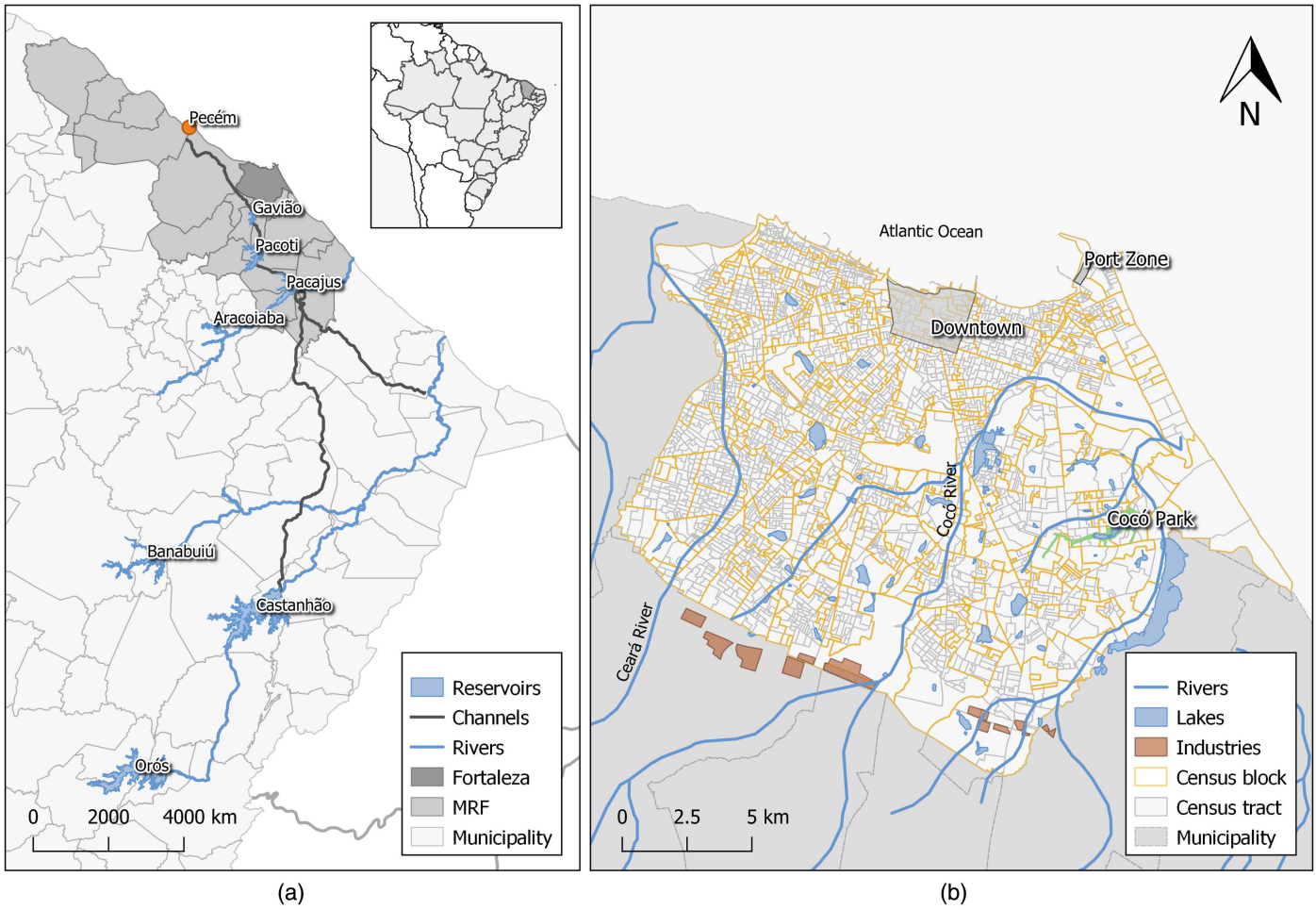

Fig. 1. (a) Jaguaribe-Metropolitano supply system; and (b) Fortaleza's census tracts and census blocks.

Fortaleza had an uneven development process, especially because the investments in urban household infrastructure and public facilities did not follow the population growth. Whereas the wealthiest areas received great improvements (e.g., pavement and electrical lines) during the 1920s and the 1930s, the urban planners gave little attention to the poorer areas of the city (IPLANFOR 2015).

In the past, recurrent and severe droughts induced human migration to Fortaleza. During the events of 1915, 1932, and 1942, the federal government installed refugee camps in the suburban areas to prevent migrants reaching the capital (Souza and Neves 2002). Today, these areas correspond to neighborhoods with high population density and subnormal agglomerate concentration, such as Pirambú, one of Latin America's largest favela communities (Garmany 2011).

Furthermore, public services and employment opportunities are concentrated in a few neighborhoods (central and eastern zones), regions that have the highest per capita income rates of the capital. Peripheral areas (western and south zones), on the other hand, lack basic services, such as sewage and garbage collection, and job opportunities. The strong spatial concentration of income in Fortaleza aggravates the urban violence rates and social tensions (IPLANFOR 2015).

Data

This research is a cross-sectional study that compares two spatial levels of aggregation with different data availability: census tracts ($n = 2{,}952$), and census blocks ($n = 182$). The data set of the CT level included only 2 input variables [average per capita income, and the Human Development Index (HDI)], whereas the data set of the CB level included 18 variables (Table 1).

CAGECE provided a data set of monthly water consumption over the year of 2010 for a total of 878,992 households. Data were provided with a household identifier, and thus could be aggregated spatially by census tracts and census blocks. The dependent variable was average daily per capita consumption for 2010, because explanatory variables were obtained for this year. We calculated average daily per capita consumption by averaging monthly household water consumption in 2010 and dividing this by the population in the census tracts and census blocks. Average daily per capita water demand in the census tracts is presented in Fig. 2.

The explanatory variables were obtained from the 2010 census conducted by the Brazilian Institute of Geography and Statistics (IBGE 2010). The 2010 census collected extensive sociodemographic information of households—grouped into census tracts—from more than 5,000 municipalities in Brazil.

At the census tract level, publicly available data are restricted to household composition and per capita income. Household composition intentionally was excluded from the CT data set because this model is meant to assess only socioeconomic aspects of the users. Instead, we included the neighborhood HDI, calculated by the Economic Development Secretariat of Fortaleza. The index is based on the 2010 census and is the geometric mean of three indicators: average monthly income of population aged 10 years or older (income), percentage of the literate population aged 10 years or older (education), and percentage of the population over 64 years old living in the neighborhood (longevity).

Table 1. Explanatory variables at CB level

ID	Variable	Unit	Mean	Standard deviation
Census tracts ($n = 2{,}952$)				
HDI	Human development index	N/D	0.362	0.194
Av. per capita income	Average per capita income	R$	2,151.15	2,424.35
Census blocks ($n = 182$)				
Demographic variables				
% female	Female residents	%	53.32	1.82
% 65+	65 years old or older	%	6.59	2.80
% 1–14	1–14 years old	%	20.74	4.71
Dem. density	Demographic density	Hab/km^2	14,451.05	8,617.47
Life expect.	Life expectancy	Years	75.25	3.53
Education				
Exp. years of schooling	Expected years of schooling	Years	10.57	0.84
% 25 + w/elem. school	25 years or older who have completed elementary school	%	62.65	15.61
% 25 + w/high school	25 years or older who have completed high school	%	46.13	18.72
% 25 + w/college	25 years or older who have completed college	%	12.95	13.27
Income				
Av. per capita income	Average per capita income	R$	830.70	728.35
% pop living in poverty	Population living in poverty	%	11.01	7.91
% pop vuln. poverty	Population vulnerable to poverty	%	30.54	16.90
Basic services for adequate housing				
% pop w/bath. & runn. water	Population living in households with bathrooms and running water	%	95.35	2.83
% pop w/garbage coll.	Population living in urban households with a garbage collection service	%	98.60	1.96
% pop w/poor water & san. services	People in households with inadequate water supply and sanitation facilities	%	1.05	0.88
Employment and vulnerability				
% 18+ econ. active	Economically active population aged 18 or older	%	49.02	4.53
% pop vuln. poverty + no elem. education	People in households vulnerable to poverty in which no one has completed elementary school	%	8.50	6.80
MHDI	Municipal Human Development Index	N/D	0.75	0.09

Note: Av. = average; Dem. = demographic; expect. = expectancy; Exp. = expected, elem. = elementary; vuln. = vulnerable; bath. = bathroom; runn. = running; coll. = collection; san. = sanitation; econ. = economically; and Hab = Habitants.

Detailed census data are released only on an aggregate level, for geographic units containing at least 400 households. Census blocks aggregate contiguous census tracts and are available for 23 Brazilian metropolitan areas (PNUD, IPEA, and FJP 2014). More than 200 indexes are provided at this level, related to aspects of demography, education, income, employment, housing, and vulnerability. Most of the indexes are classified by sex and age; thus, to reduce the number of variables, some of them were merged. The final data set included the potentially relevant variables of each category, reducing the indexes to 18 variables expected (Table 1). Variables were chosen to assess socioeconomic inequalities and to explain consumer behavior.

Demographic variables initially included 85 indexes, which were narrowed to 5, assessing household composition; population distribution across the city; and environmental health, represented by life expectancy (Gulis 2000). The percentage of male residents was excluded because it perfectly correlated with percentage of female residents (Pearson correlation coefficient = 1) and would not add information to the model.

Variables related to education assess different stages of formal learning. The Brazilian education system is divided into two levels: basic and higher education. Basic education includes three stages: preschool (for children 0–5 years old), elementary school (for children 6–14 years old), and high school or secondary education (for children 15–17 years old). A high school diploma is mandatory for admission to higher education.

The category of income included three variables. Those considered as living in poverty had a per capita household income equal to or less than one-fourth the minimum wage, whereas those vulnerable to poverty had a per capita household income of less than one-half the minimum wage. These variables were included because average per capita income alone could disguise information about the income gap. Variables regarding basic services for adequate housing reflected the health condition of the inhabitants (Montgomery and Elimelech 2007).

In the category of employment and vulnerability, the percentage of the economically active population aged 18 or older accounted for people in the job market or trying to join it. The Municipal HDI (MHDI) reflected the three dimensions of the global HDI: longevity, education, and income. HDI-longevity is measured by life expectancy at birth. HDI-education is the geometric mean of two indicators: the education of the adult population (Weight 1), and the school flow of young population (Weight 2). HDI-income is the municipal per capita income, including those who do not have any income.

Pearson's parametric correlation coefficient was used to estimate the association between per capita water consumption and the independent variables and to further analyze the ranking provided by the RF (Fig. S1). Other than garbage collection service, households with inadequate water supply and sanitation, and demographic density, all variables were associated strongly with water consumption. Independent variables also correlated with each other, such as per capita income, which was associated with life expectancy at birth ($r = 0.74$), percentage of college educated people ($r = 0.94$), and MHDI ($r = 0.81$). Correlated variables usually are avoided because they might contain redundant information, but a high correlation does not mean a lack of variable complementarity

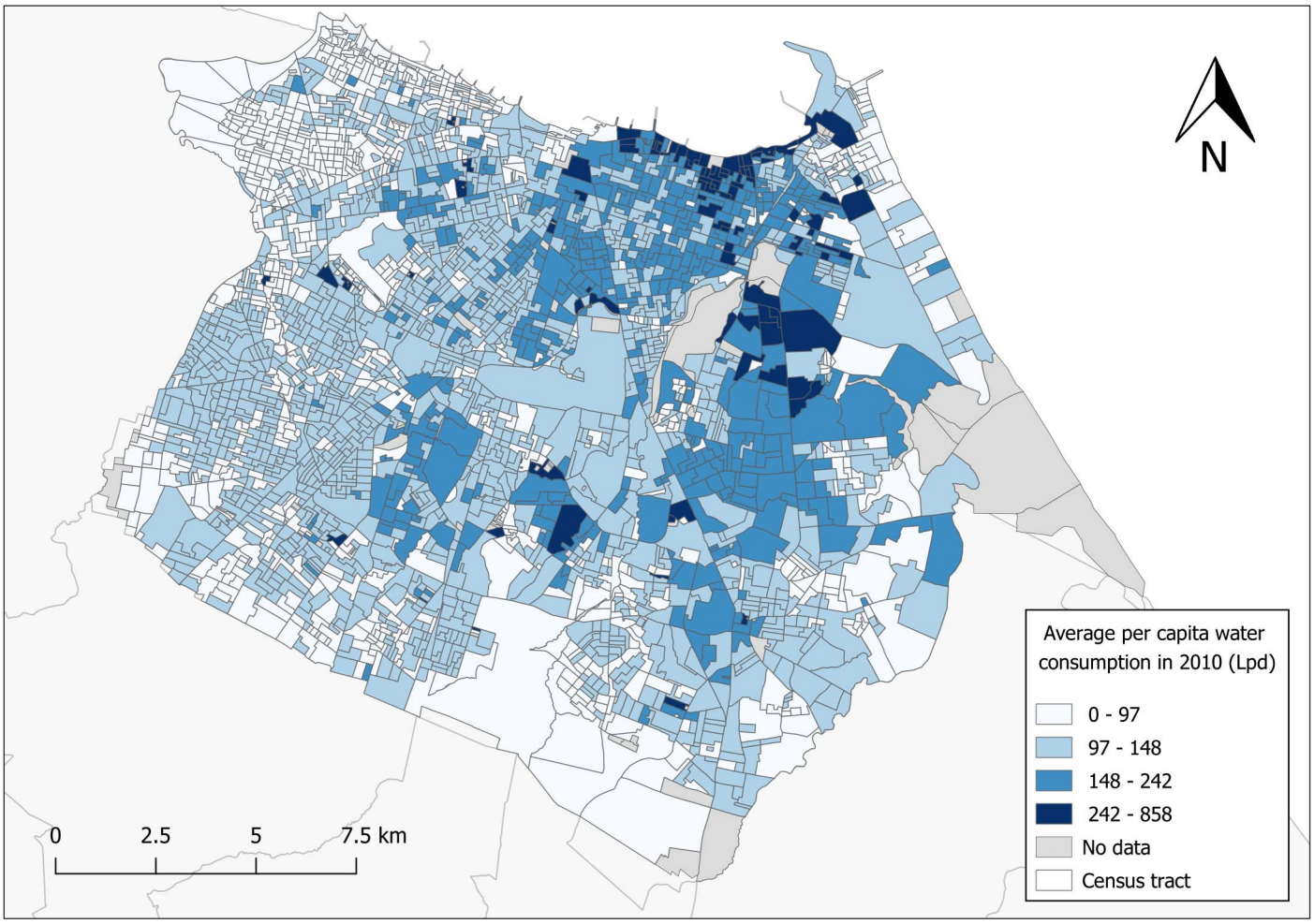

Fig. 2. Average daily water consumption in 2010 in the census tracts of Fortaleza. (Data from CAGECE, unpublished data, 2010.)

(Guyon and Elisseeff 2003). These variables were maintained because the initial intention of ranking the variables was to understand the relationship between them and to find a reduced group of variables that could explain water demand for clustering. In addition, the RF method is appropriate for dealing with correlation (further explanation is provided subsequently). However, when selecting the input variables for the predictive model, the IIS method was used to avoid redundant information.

Methods

The methodology of this study was divided into three sections (Fig. 3): (1) variable importance using a RF; (2) clustering and spatial analysis of demand and sociodemographic characteristics with a SOM; and (3) variable selection with the IIS method and predictive model using ANN.

The first part of this study investigated which sociodemographic characteristics drive consumer behavior and water consumption. This analysis was performed at the census block level, which had 18 explanatory variables. A RF was used to define variable importance and to study the relationship between them. After defining the most relevant sociodemographic variables driving water demand, a SOM was used to cluster data and to visualize the spatial patterns present in these variables. The census tract data also were clustered, to compare spatial-level aggregation. The predictive model was built using an ANN, and it was tested for both spatial levels, CB and CT. The first considered the variables iteratively selected with the RF and ANN models, whereas the last had only two explanatory variables.

Algorithms and Model Specifications

Random Forest

A random forest (Breiman 2001) is a supervised learning algorithm mainly used for regression and classification tasks. A RF is based on the combination of many classification and regression tree models trained with bootstrapping aggregation. The combined result of many decision trees is used for prediction. The general steps in constructing a random forest are (Hastie et al. 2009)

1. Draw a bootstrap sample of size a_n from the original data set. These observations are used to build the tree.
2. Grow a tree (T_b) to the bootstrapped sample by recursively repeating the following steps for each terminal node of the tree until the minimum node size (nodesize) is reached
   a. Select a subset of variables at random among the original variables. The number of variables to be drawn is denoted mtry.
   b. Pick the best variable/split point among the selected variables.
   c. Split the node into two daughter nodes.
3. Summarize over all trees. For classification trees, use the majority vote of the classes predicted by the trees. For regression trees, use the average (Hastie et al. 2009)

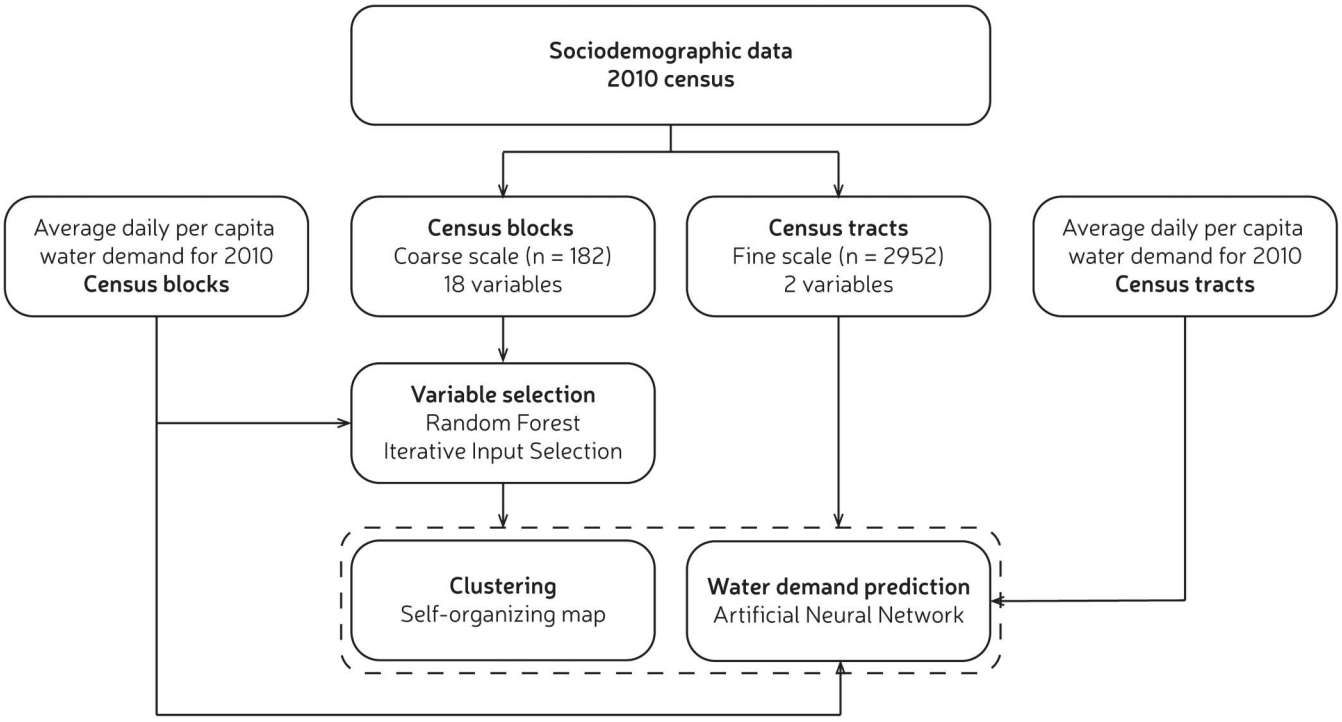

Fig. 3. Methodological steps.

$$y(\mathbf{x}_i) = \hat{f}_{RF}^{N}(\mathbf{x}_i) = \frac{1}{N}\sum_{b=1}^{N} T_b(\mathbf{x}_i) \tag{1}$$

where $\mathbf{x}_i$ = vector of independent variables; $T_b(\mathbf{x}_i)$ = single regression tree grown by bootstrapped samples and a subset of variables; and N = number of regression trees.

An important feature of random forests is the use of out-of-bag samples (Hastie et al. 2009). The training set of each tree is selected using a bootstrap, and the observations left out by the bootstrap sampling are the out-of-bag sample. This sample is used for performance evaluation, providing an unbiased estimate of the prediction error (Genuer et al. 2010).

RFs are efficient and widely used for variable selection and prediction. They are applicable to problems with nonlinear relationships between the variables, and effectively can handle small sample sizes (Biau and Scornet 2016). The tree-building process of random forests implicitly allows for interaction and high correlation between features (Ziegler and König 2014). Although variable importance decreases when highly correlated variables are added to a RF model, the relative position between the variables is preserved (Genuer et al. 2010).

After growing each regression tree, the out-of-bag sample is passed down the tree and the mean squared error (MSE) is computed. To assess the importance of a specific predictor variable, its values are permuted randomly for the out-of-bag sample, and the MSE is computed again. The increase in the MSE (IncMSE) resulting from the permuting is averaged over all trees and is used to measure the variable importance. Therefore, if a predictor is important for the model, randomly assigning other values for that variable should have a negative influence on prediction.

The IncMSE was used to rank the variables. Different criteria were defined for variable selection: for clustering, 45% of the least important variables were removed; for prediction, the IIS method was performed.

The model was validated through leave-one-out cross-validation to reduce bias in training data. In this approach, one data point is used for validation, and the training set is composed of $n - 1$ samples, where n is the number of observations. The final error estimate is based on the average of the results of all n tests (Witten and Frank 2016); for this study, the error estimate was based on the average IncMSE for 182 tests. To obtain a stable solution and to assess the variance of the measures, 100 runs of the model were performed, and the median of the mean IncMSE was used to rank the variables. Further details about the parameters of the RF model are presented in Supplemental Materials Text S1.

Accumulated Local Effect

To assess the main effects of the individual predictor variables, they were visualized with accumulated local effect (ALE) plots (Apley and Zhu 2016). ALE plots describe how variables influence the prediction of a machine learning model on average, and are appropriate for highly correlated inputs (Molnar 2019). To estimate local effects, the variable is divided into many intervals and the differences in the predictions are computed. The grid that defines the intervals consists of the quantiles of the variable distribution, to ensure that each interval contains the same number of observations. The uncentered effect for each variable is estimated as follows (Molnar 2019):

$$\hat{\tilde{f}}_{j,\mathrm{ALE}}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum i{:}x_j^{(i)} \in N_j(k) \left[f\left(z_{k,j}, x_{Jay}^{(i)}\right) - f\left(z_{k-1,j}, x_{Jay}^{(i)}\right) \right] \tag{2}$$

where k = number of intervals of variable x; n = number of observations in interval k; N = neighborhood, i.e., observations within an interval; z = grid value; x = variable of interest; and f = predictive function. This effect is centered so that the mean effect is zero

$$\hat{f}_{j,\mathrm{ALE}}(x) = \hat{\tilde{f}}_{j,\mathrm{ALE}}(x) - \frac{1}{n}\sum_{i=1}^{n}\hat{\tilde{f}}_{j,\mathrm{ALE}}(x_j^{(i)}) \quad (3)$$

The value of the ALE represents how much the output of the model deviates from the average prediction at a certain value of the variable of interest.

Self-Organizing Map

A self-organizing map clusters high-dimensional data vectors and reduces them to a one- or two-dimensional map (Kohonen 1982). The lattice of the grid can be either hexagonal or rectangular, but hexagonal is better for visualization (Vesanto and Alhoniemi 2000). The typical structure of a SOM consists of an input layer and an output layer. The input layer contains one neuron for each variable in the data set. The neurons in the output layer are connected to the input neurons through adjustable weights; each neuron i has a weight vector $\boldsymbol{w} = (w_{i1}, w_{i2}, \ldots, w_{id})$, where d is the dimension of the input space. These neurons relate to their neighbors according to topological connections, i.e., the map is neighborhood preserving. The general steps in the learning algorithm of the self-organized map are (Chaudhary et al. 2014)

1. Initialize the weight vectors $\mathbf{w}_i$s of the $m \times n$ neurons.
2. Randomly select an input vector $\mathbf{x}(t)$, which represents the pattern that is presented to the neurons in the output layer.
3. Find the winner neuron $\boldsymbol{c}$ or the best matching unit based on the minimum distance Euclidean criterion

$$\boldsymbol{c} = \mathrm{argmin}||\mathbf{w}_{\boldsymbol{i}}(t) - \mathbf{x}(t)|| \quad (4)$$

   where ||.|| = Euclidean distance measure; and $\mathbf{x}(t)$ and $\mathbf{w}_i(t)$ = input and weight vectors of neuron at iteration t, respectively.
4. Update the weight vector of the neurons using the following equation:

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + h_{c,i}(t)[\mathbf{x}(t) - \mathbf{w}_i(t)] \quad (5)$$

   where $h_{c,i}(t)$ = Gaussian neighborhood function

$$h_{c,i}(t) = \alpha(t) \times \exp\left(-\frac{||r_c - r_i||^2}{2\sigma^2(t)}\right) \quad (6)$$

   where r = coordinate position of neuron on map; $\alpha(t)$ = learning rate; and $\sigma(t)$ = neighborhood radius. Both $\alpha(t)$ and $\sigma(t)$ decrease monotonically. For all the input data, repeat Steps 2–4.

The main parameters of the SOM are the grid size, the training rate, and the neighborhood size. There is no theoretical justification in the literature for choosing the optimal grid size of the output layer. Previous studies used different criteria (Kalteh et al. 2008), but the general recommendation is to define the size by trial-and-error (Kohonen 2014). The map quality can be evaluated through the resolution of the cluster structures and the node counts, i.e., how many samples are mapped to each output neuron. An ideal map size does not have areas with large values or many empty nodes. A 6×6 network (CB level; coarse scale) and a 12×12 network (CT level; fine scale) were considered the most suitable for the problem. Larger maps resulted in many empty nodes and/or fewer than two data points per node. Further details of the parameters of the model are presented in the Supplemental Materials Text S2.

Cluster Validation

Two cluster validity measures were used to choose the best number of clusters: the Dunn index, and the silhouette index. The Dunn index (Dunn 1974) is the minimum distance between observations in different clusters divided by the largest intracluster distance. A higher Dunn index indicates better clustering and smaller cluster sizes. It is computed as

$$\mathrm{DI} = \frac{\min_{1\le i<\le m}\delta(C_i, C_j)}{\max_{1\le k\le m}\mathrm{diam}_{C_k}} \quad (7)$$

where m = number of clusters; $\delta(C_i, C_j)$ = dissimilarity function between clusters Ci and Cj; and diam_{Ck} = diameter of a cluster C_k. The dissimilarity function is defined as

$$\delta(C_i, C_j) = \min_{x\in C_i, y\in C_j} d(x, y) \quad (8)$$

where d = Euclidean distance. The diameter of a cluster C is defined as the Euclidean distance between the two farthest points inside the cluster.

The silhouette index (Rousseeuw 1987) is given by

$$S_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \quad (9)$$

where a = mean Euclidian distance between an observation and all other data points in the same cluster; and b = mean Euclidian distance between an observation and all other points in the next cluster.

The silhouette coefficient is the mean of all samples in the data set, and it reveals the capability of clustering similar objects in a group and minimizing interclass dissimilarity. The values range from -1 to 1, where $S = 1$ corresponds to a high quality of clustering, and $S = -1$ corresponds to false clustering. The silhouette plot shows how close each point in one cluster is to points in the neighboring clusters.

The clusters also were identified through a graphical method based on the unified distance matrix (U-matrix), which shows the Euclidean distance between output nodes of neighboring map units.

Artificial Neural Network

Artificial neural networks are statistical models built through an iterative self-learning process. An ANN is a network of weighted connections between neurons (nodes). The weights are defined during the training process, and are updated according to the chosen algorithm. A network is composed of at least two layers: input and output. A multilayer perceptron (MLP) network has at least one hidden layer in addition to the input and output layers, with a nonlinear activation function. The general equation for an MLP is (Bishop 1995)

$$\mathbf{y}_k = f_{\mathrm{outer}}\left[\sum_{j=1}^{M}\mathbf{w}_{kj}^{(2)} f_{\mathrm{inner}}\left[\sum_{i=1}^{d}\mathbf{w}_{ji}^{(1)}\mathbf{x}_i + \mathbf{w}_{j0}^{(1)}\right] + \mathbf{w}_{k0}^{(2)}\right] \quad (10)$$

where $\mathbf{y}_k = k$th output; f_{outer} = output layer transfer function; f_{inner} = input layer transfer function; $\mathbf{w}$ = weights and biases; and $i = i$th layer.

The domestic water demand was projected with a multilayer perceptron network and trained with a back-propagation algorithm (Rumelhart et al. 1986). Back-propagation is a supervised learning method that adjusts the weights by minimizing the error between the model output and the observed values. Determining the number of hidden layers is a difficult task, and there is no general rule for doing so (Reed and Marksii 1999), but one or two hidden layers usually are enough to solve any nonlinear problem (Lippmann 1987). A multilayer perceptron with one hidden layer was used in this study. Adding more hidden layers would increase not only

computational time, but also the number of parameters, and a larger training data set would be necessary.

At the census block level, the input variables were defined using the IIS method. At the CT level, a k-fold cross-validation analysis was conducted. In this approach, the data set is divided into k subsets: $k-1$ subsets are used to train the model, and the remaining subset is used for testing. This process is repeated until all k subsets are used for testing; then the average and standard deviation performance are computed. This study used $k = 5$.Because variables were in different scales and units, data were normalized by min-max scaling. The parameters used for performance evaluation were mean absolute error (MAE), RMS error (RMSE), and R-squared (R^2).

Iterative Input Selection

The IIS method, proposed by Galelli and Castelletti (2013), is a tree-based method for the selection of inputs with minimum redundancy, while keeping the most significant variables for prediction. This study adapted the IIS approach to incorporate the RF and the ANN models.

The algorithm is divided in three steps (Galelli and Castelletti 2013): (1) the IIS algorithm runs an input ranking algorithm to sort the variables with a nonlinear statistical measure of significance; (2) the first p variables in the ranking are individually used as the input to a model-building algorithm, so that p single-input–single-output (SISO) models are constructed, and their performance is evaluated with a suitable metric, and the best-performing model is added to the final selection of input variables; and (3) the selected variables are used as an input to the model-building algorithm [multi-input–single-output (MISO) model], and the residuals are calculated.

The residuals are used as the output variable in the first two steps to ensure that the next selected variable will not contain redundant information. These steps are iterated until either a repeated variable is selected in Step 2 or the performance of the SISO model does not improve significantly. The minimum improvement in significance is defined by the parameter ε.

At each step, both the SISO and MISO models are evaluated with a k-fold cross-validation approach. In this study, the metric for evaluating model performance was the R-squared. Although the original IIS approach uses a model-free input ranking algorithm, here, the RF model was chosen, with IncMSE as the significance measure, to be consistent with the first step of the methodology. The parameters for the RF were the same as those described in section "Methods." Although this strategy might slow the algorithm, it

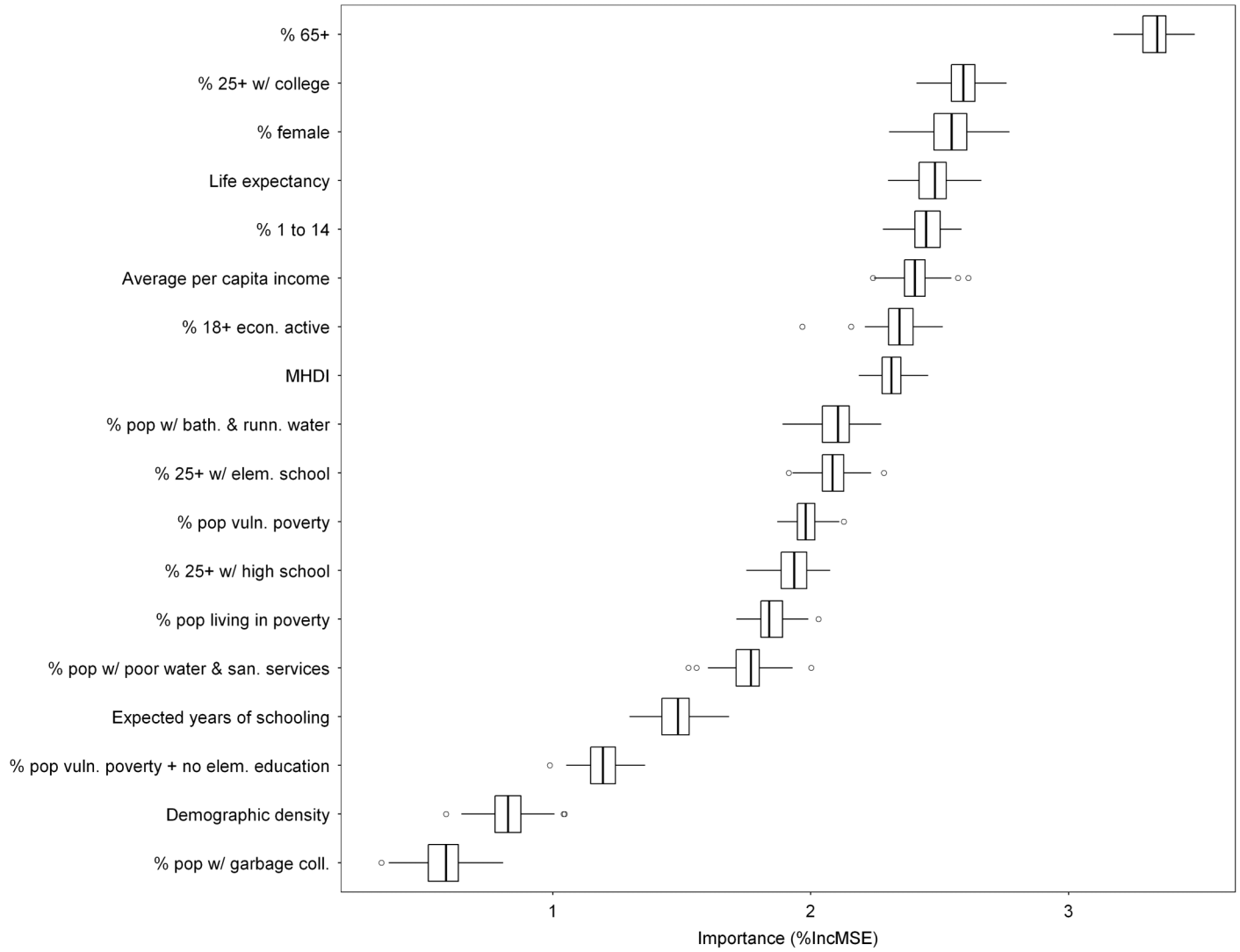

Fig. 4. Variable importance according to RF. Boxplots represent the variation in the average %IncMSE for 100 runs of the model. Variables are ranked according to the median value of the importance measure. Table 1 describes the explanatory variables.

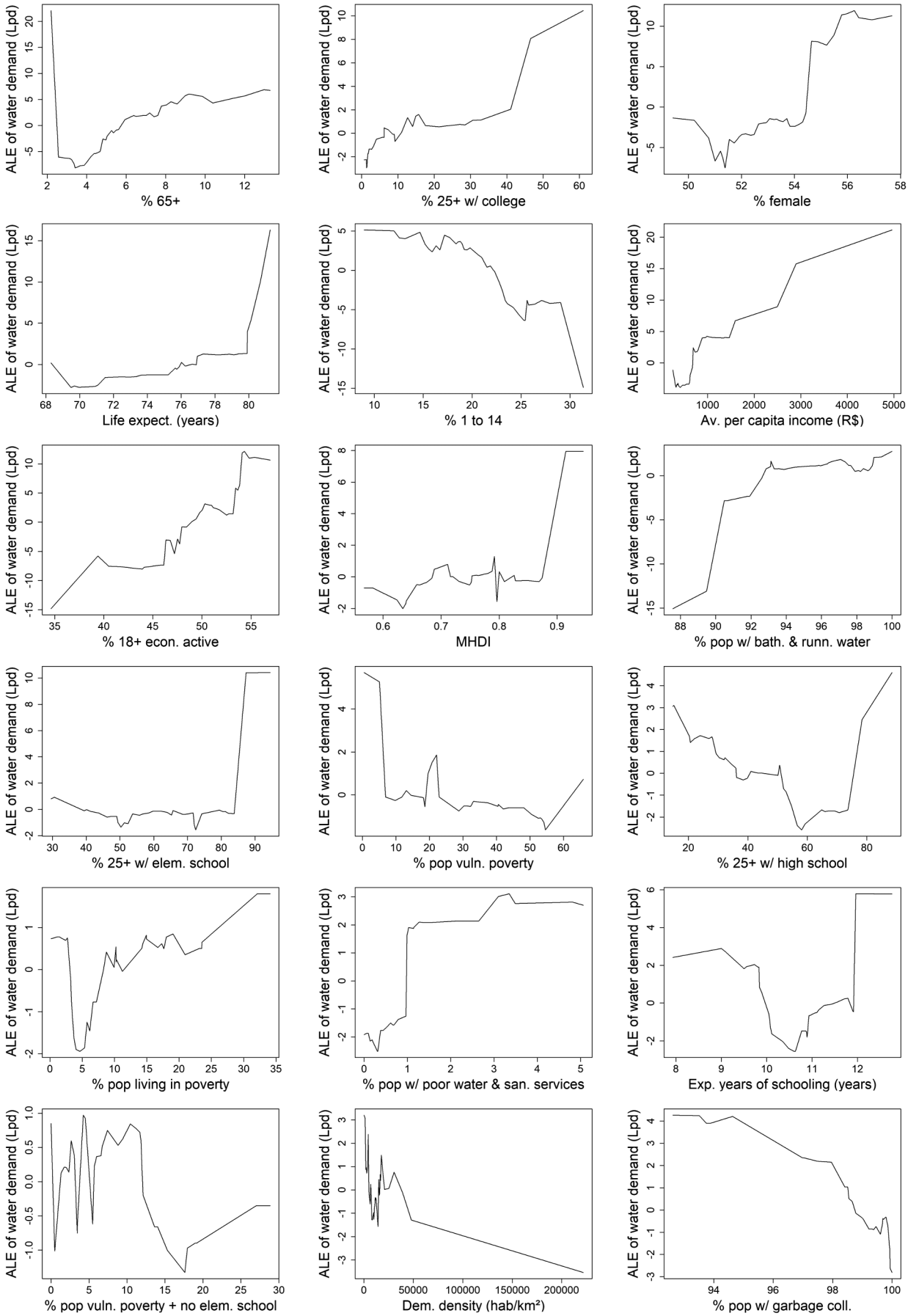

Fig. 5. Accumulated local effect plots for the RF model.

still provides the desired ability to detect nonlinear relationships and to handle variables with different dimensionality. The model-building algorithm was the ANN, with the parameters previously mentioned. A sensitivity analysis was performed to choose the IIS method parameters. The number of SISO models evaluated at each iteration p was set to 1, 5 and 10; the number k for the cross-validation was 2, 5, and 10; and ε varied between 0 and 0.1, with an incremental value of 10^{-2}.

Results and Discussion

Variable Importance

The variables were ranked according to the median of the increase in MSE for 100 runs of the RF model (Fig. 4). The interquartile range was small (less than 0.1) for all the variables, indicating that the importance measure was stable. The median importance ranged from 2.58 to 2.31 between the second and the eighth variables, meaning that the relative position among them was irrelevant for model interpretation.

Variables that assess household composition (percentage of elderly and women) and education (percentage of residents with college degree) were the most relevant to water demand prediction in Fortaleza. Life expectancy, percentage of children, and average income also were of high importance. Variables with low correlation ($r < 0.2$) to water demand, such as garbage collection coverage, had low importance scores. Some highly intercorrelated variables ($r > 0.7$) were ranked at the top, e.g., percentage age 65+ and percentage of females, percentage age 1–14 and percentage age 25+ with college education, and percentage age 25+ with college education and life expectancy.

The significance of household composition for water demand forecasting was corroborated by several studies (House-Peters et al. 2010; Bennett et al. 2013; Matos et al. 2014; Hussien et al. 2016; Villarin and Rodriguez-Galiano 2019). Life expectancy and the presence of indoor bathrooms and running water might be useful to assess quality of life. Furthermore, the latter has a direct relationship with water demand.

The accumulated local effect plots were helpful to interpret the effect of the explanatory variables on the average prediction of water demand (Fig. 5). The average per capita income had a strong positive effect on the prediction. The influence of income in water use has been explored extensively in other studies (House-Peters et al. 2010; Shandas and Parandvash 2010; Liu et al. 2015; Villarin and Rodriguez-Galiano 2019). Households with higher income are more likely to install water-saving devices and water storage units, e.g., cisterns and water tanks (Grande et al. 2016). Although it would be expected that these mechanisms would reduce household consumption, past studies led to divergent conclusions (Olmstead and Stavins 2009). High-income households are less likely to be concerned about saving water than are low- and medium-income households, who tend to maintain a lower consumption to avoid water shortage.

Percentages of children and elderly had opposite effects on water demand. The average prediction increased with increasing percentage of elderly (when above 4%), but decreased with increasing percentage of children. An inverse relationship between households with children and water demand also was found in previous studies (Schleich and Hillenbrand 2009; Hussien et al. 2016). However, different consumption patterns were detected in Spain (Martinez-Espiñeira 2002), Portugal (Matos et al. 2014), and Italy (Musolesi and Nosvelli 2007), where water use tends to decrease with age. A positive relationship between percentage of elderly and the predictions could imply an increase in water demand in the next 20 years, because a demographic trend of population aging is expected in Fortaleza (Barreto and Menezes 2014).

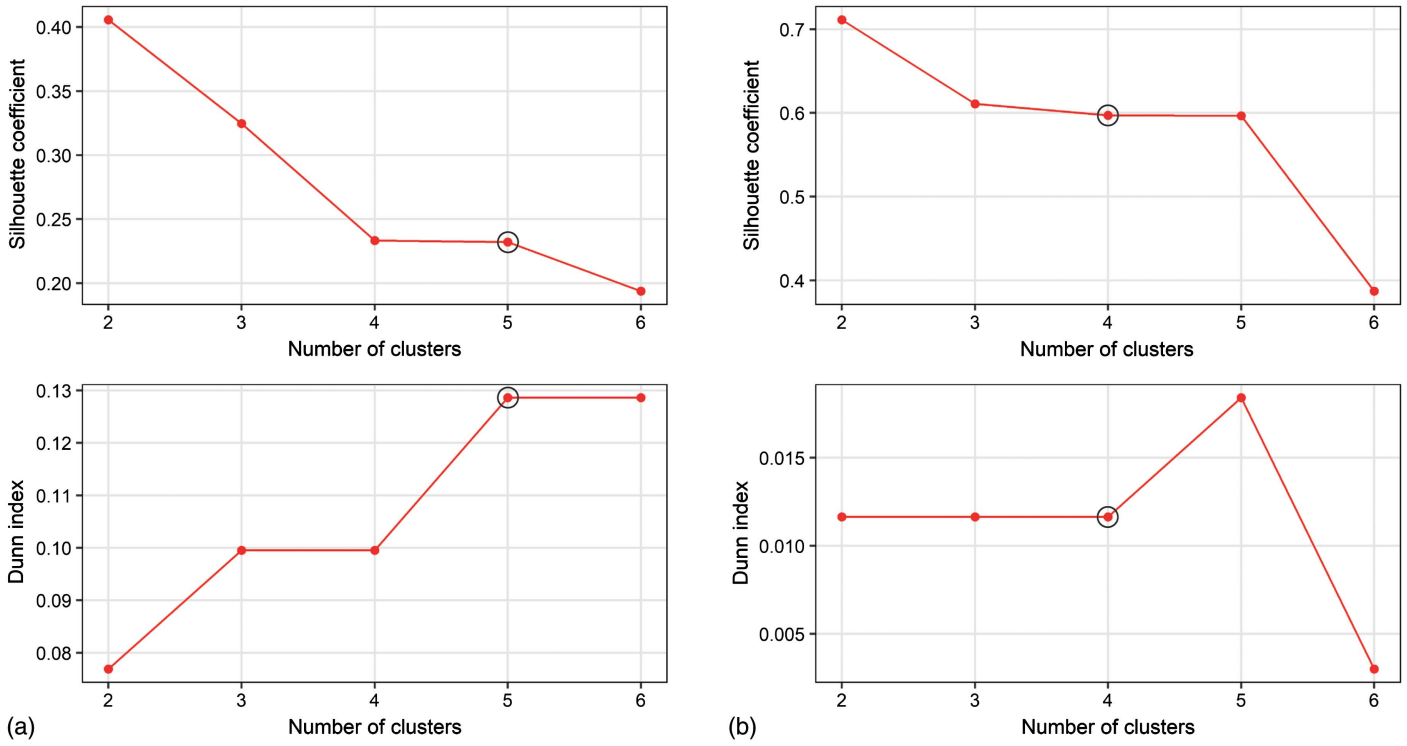

Fig. 6. Dunn index and silhouette index for different number of clusters at the (a) census block level; and (b) census tract level. The chosen number of clusters for each model are circled.

J. Water Resour. Plann. Manage., 2021, 147(1): 05020026

Some of the variables had a very significant effect on prediction after reaching a threshold, such as percentage of females, life expectancy, MHDI, and percentage of adults who completed college and elementary school. The effect of the presence of bathrooms and running water in the households on average water demand was more significant when the percentage ranged between 88% and 93%.

The variables which decreased in the RF ranking had little effect on the prediction. An increase in garbage collection coverage from 96% to 98%, for example, reduced average per capita water demand by only one unit. Some of these predictors had a complex relationship with the outcome and were difficult to interpret, such as years of schooling, percentage of population vulnerable to poverty + no elementary education, and demographic density.

Spatial Analysis of Water Demand

After removing 45% of the least important variables from the ranking provided by the RF, the 10 remaining sociodemographic variables at the CB level were used to cluster water demand using the

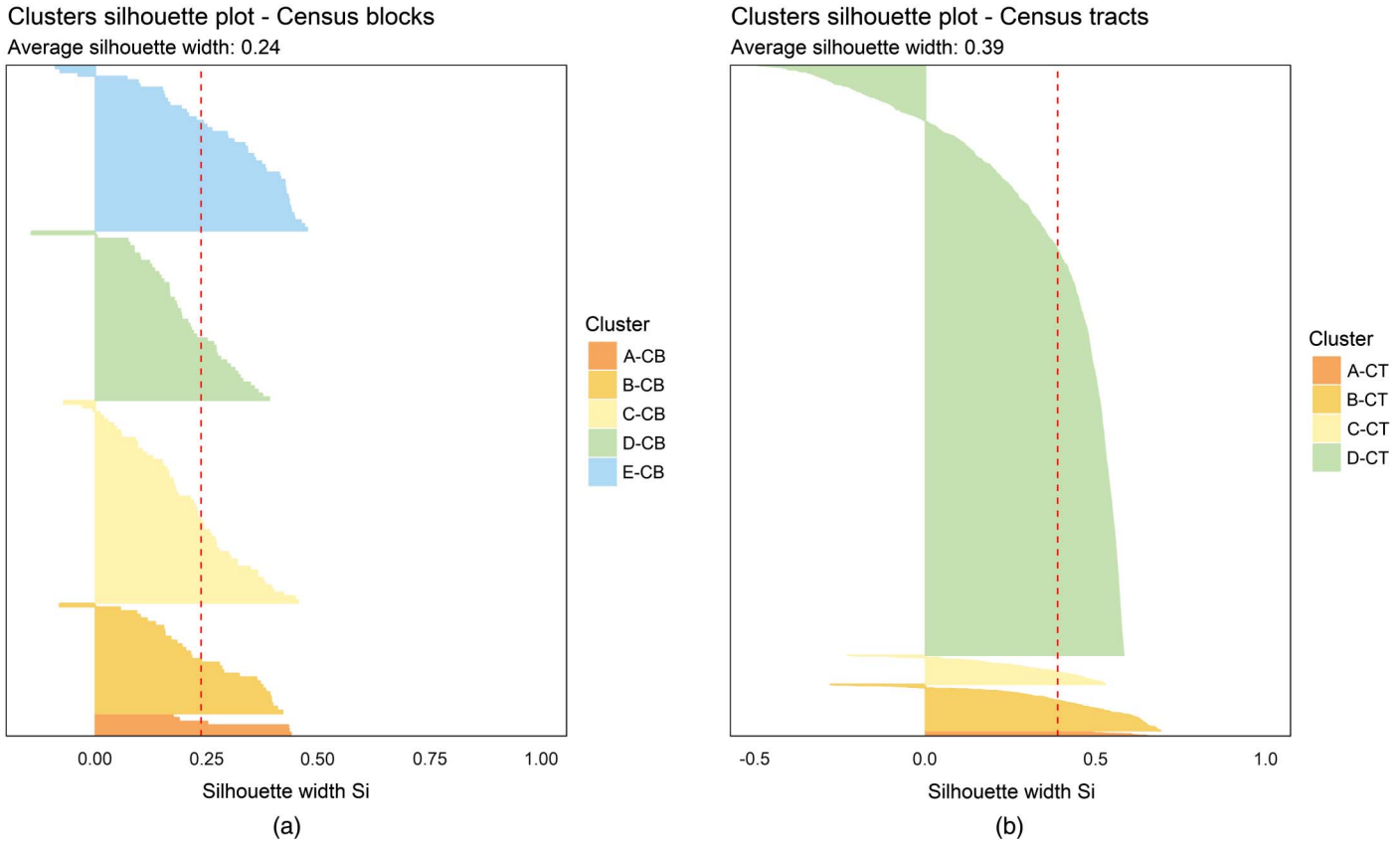

Fig. 7. Clusters silhouette plot for (a) census blocks; and (b) census tracts aggregation. For each census block or census tract, a straight horizontal line denotes the silhouette coefficient. Each object is shaded according to the correspondent cluster, and the dashed line represents the average silhouette width.

Table 2. Characteristics of SOM clusters defined using 10 most important explanatory variables at census block level

	Cluster				
Variables	A-CB ($n = 6$)	B-CB ($n = 30$)	C-CB ($n = 55$)	D-CB ($n = 46$)	E-CB ($n = 45$)
Total area (km^2)	16,655,727	46,821,907	72,779,209	60,424,798	101,088,280
Total population	127,415	298,058	637,127	496,293	767,285
Average water demand (L/day)	204.76	135.25	126.52	107.43	105.00
PELD (%)	9.89	9.84	7.18	5.70	4.18
COLL (%)	51.95	31.44	14.00	5.20	2.05
PFEM (%)	55.82	55.72	53.73	52.63	51.57
LIFEXP (years)	80.89	79.67	77.02	73.85	70.79
P1T14 (%)	14.29	14.57	18.62	22.60	26.38
APCI (R$)	3,622.77	1,593.81	803.31	479.12	342.57
P18EAP (%)	54.93	53.76	51.25	47.90	43.48
MHDI	0.925	0.860	0.786	0.708	0.643
BTHRW (%)	98.23	96.37	96.97	95.11	92.56
ELSCH (%)	89.67	82.08	70.32	55.73	43.81

Note: Except for area and population, variables are represented by mean value for all census blocks in each cluster. PELD = 65 years old or older; COLL = 25 years or older who have completed college; PFEM = female residents; LIFEXP = life expectancy; P1T14 = 1–14 years old; ACPI = average per capita income; P18EAP = economically active population aged 18 or older; BTHRW = population living in households with bathrooms and running water; and ELSCH = 25 years or older who have completed elementary school.

SOM. The variables at the CT level (HDI and per capita income) also were used to create clusters.

At the CB level, the Dunn index indicated that five or six clusters were the best choice, but a larger silhouette coefficient was obtained for five clusters [Fig. 6(a)]. Although two and three clusters had larger silhouette coefficients, five clusters were preferred because this was more convenient for the analysis of Fortaleza's heterogeneities. CB data presented rather low silhouette widths (ranging between 0.2 and 0.5) [Fig. 7(a)], but the clusters were substantially different from each other, especially in percentage of females, percentage of college graduates, and average per capita income (Table 2). For example, the average per capita income in Cluster E-CB was less than 10% that of Cluster A-CB.

The SOM map for CB data and its clusters are represented in the U-matrix (Fig. 8). The heat maps in Fig. 8 shows the distribution of the explanatory variables across the SOM. They reveal a direct relationship between average per capita income, education level (percentage 25+ years of age with elementary school education and percentage 25+ years of age with college education), MDHI, and percentage of economically active population. These had an inverse relationship with percentage of children. Percentage of females and elderly also had a direct connection.

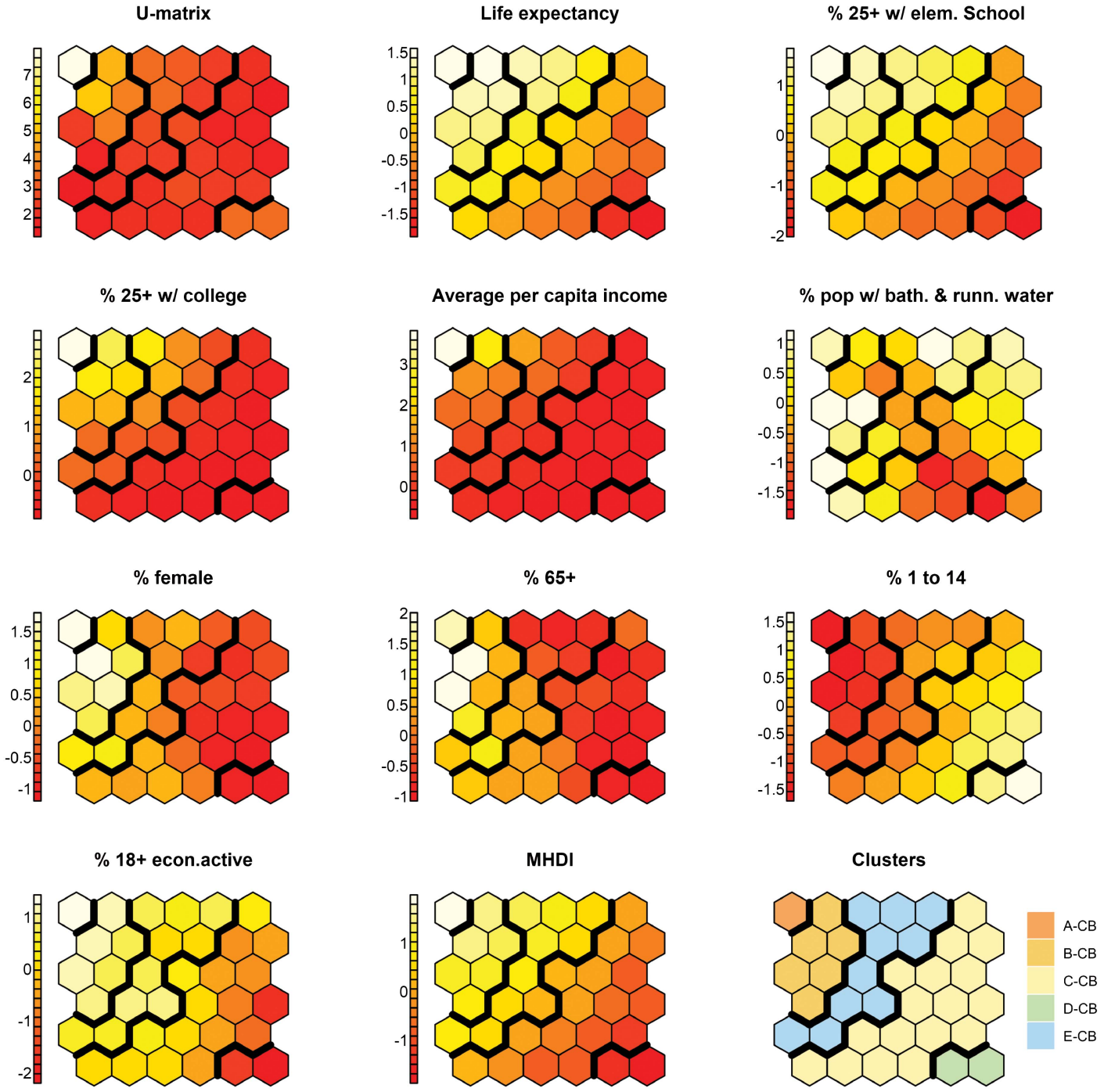

Fig. 8. SOM heat maps for explanatory variables at census block level. The gradient represents the Euclidean distance between each node and its neighbors: lighter shading indicates large distances and darker shading indicates small distances. Table 1 describes the explanatory variables.

The CB cluster's spatial distribution is represented in Fig. 9, and their characteristics are listed in Table 2. Neighborhoods with high HDI and elevated per capita income were clustered together (A-CB and B-CB). These also were the areas with the highest water consumption rates. Further comments on the cluster divisions are provided in the Supplemental Materials.

At the CT level, the silhouette coefficient indicated that two clusters would be the best choice, but three, four, or five also were acceptable [Fig. 6(b)]. The largest Dunn index was obtained for five clusters, but four clusters was considered the most suitable for further analysis. The four clusters at the CT level had moderate silhouette values, with an average width of 0.39 and some misclassified CTs (negative S_i), especially in Cluster D-CT [Fig. 7(b)]. Overall, the CT data set presented relatively good clustering.

The heat maps of the CT level SOM show a direct relationship between HDI and average per capita income (Fig. 10). The clusters were less representative than those of the CB level (Fig. 11), probably because only two variables were used to create them. Areas with elevated average per capita income and HDI were assigned to Clusters A-CT and B-CT, which also had elevated water consumption (Table 3). Census tracts with medium water consumption were clustered in C-CT. Cluster D-CT, which had almost 90% of the population, incorporated census tracts with low per capita income and water use.

To verify that clusters were a good representation of water demand patterns, the water demand in each census tract and census block was compared with the average water demand of their corresponding clusters and the relative error was calculated. The mean relative error for each cluster is presented in Table 4. CT-level clustering (finer scale) resulted in better separated clusters than did CB-level clustering (coarser scale), but was worse for water demand assessment (higher relative errors).

Although clustering could be used to improve prediction, this would have decreased the ANN performance because some clusters had very few data points (A-CB, for example, contained only six census blocks). Sociodemographic-based clustering allows the incorporation of spatial heterogeneities in economic development when projecting long-term water demand. Clustering at a fine scale with fewer variables provided better separated clusters, but the coarse scale was more convenient for urban planning and water demand estimation.

Predictive Model

The input variables for the ANN model at the census block level (ANN-CB) were chosen with the IIS method. The sensitivity analysis (Table S1) indicated that the best performing selected models were those with 5 SISO models and $k = 10$ in the cross-validation.

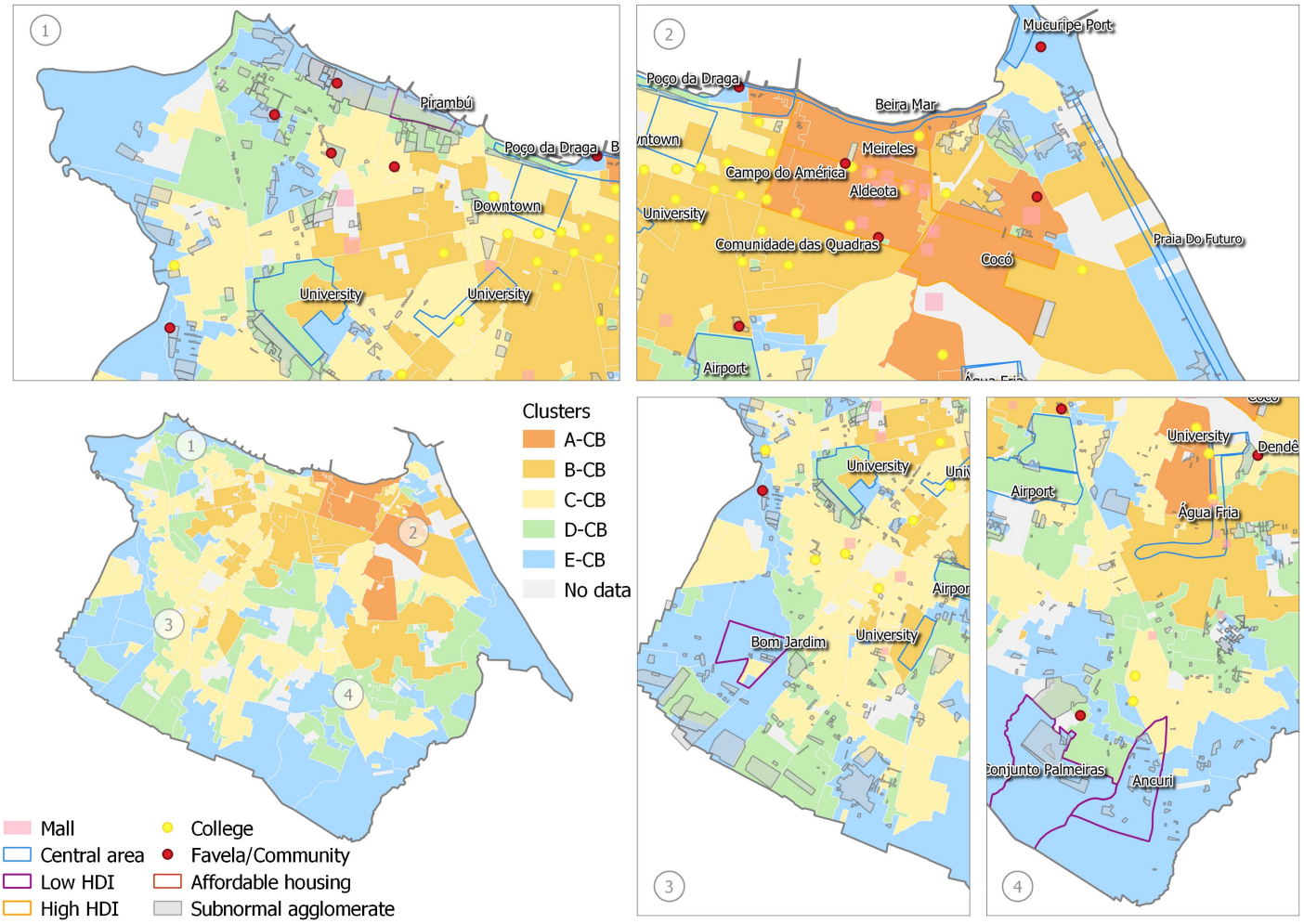

Fig. 9. Clusters on the CB level defined by the SOM using the 10 most important explanatory variables for water consumption, defined by the RF. Central areas of Fortaleza are highlighted.

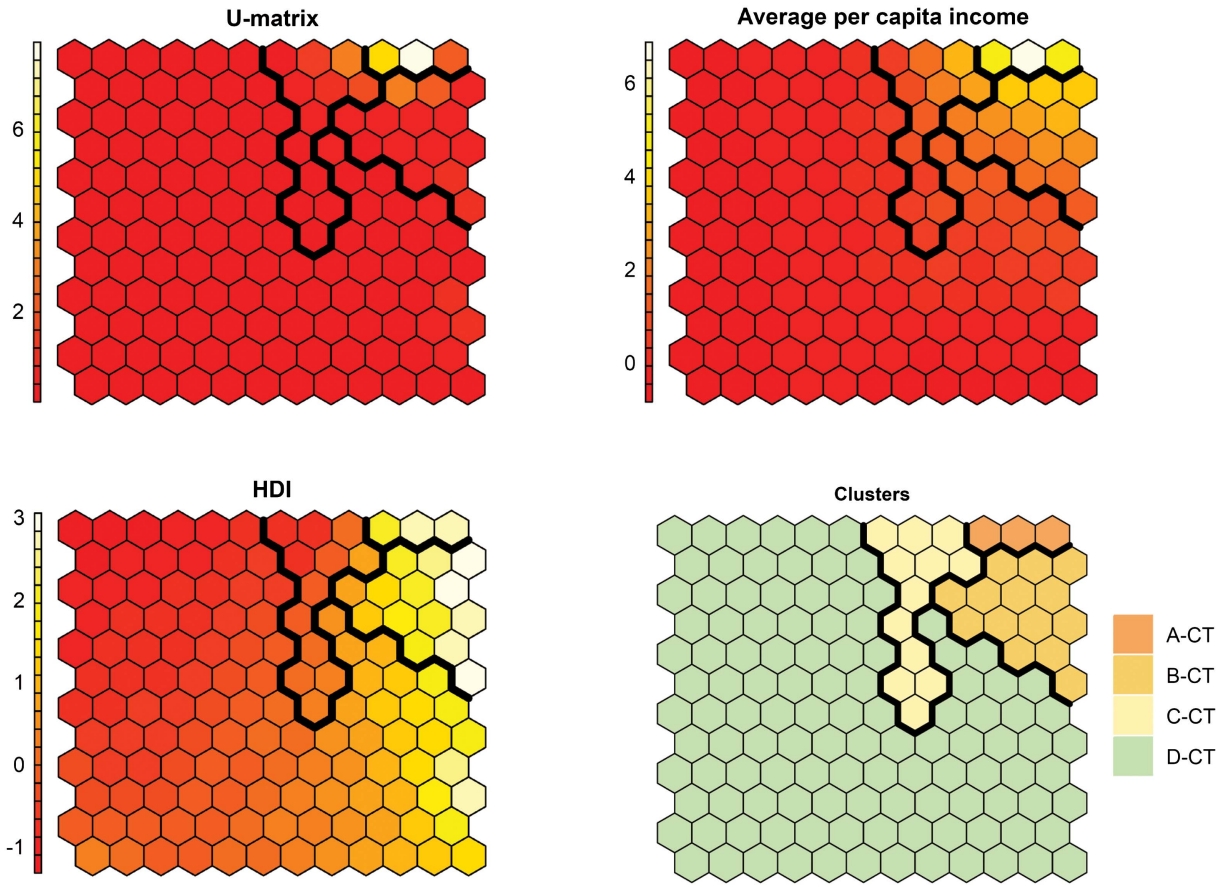

Fig. 10. SOM heat maps for explanatory variables at the CT level. Table 1 describes the explanatory variables.

The performance was similar for a tolerance ε ranging between 0 and 0.03, all providing the same number of variable inputs. In the final selection, ε was set to 0.01. The variables selected with these parameters and the model performance obtained with the inclusion of each variable are presented in Fig. 12. The first two variables selected with IIS (average per capita income and percentage age 1–14) were at the top of the RF ranking, whereas the third (percentage of population living in poverty) had a rather low score. These three variables can describe water demand in Fortaleza, with the average per capita income functioning as a proxy for socioeconomic aspects of the households, percentage age 1–14 describing demographic aspects, and percentage of population living in poverty adding information related to the vulnerability of the population.

The performance of ANN models at the CT (fine scale) and CB (coarse scale) levels is presented in Table 5. The results show that the CT model had a slightly better performance than the CB model in terms of R^2. One explanation is that the larger number of observations in the CT data set benefitted the training process of the MLP, which, as previously pointed out, requires large data sets. The ANN-CB model had only 182 observations, whereas the ANN-CT model had 2,952 and 2 independent variables.

Water-use patterns can differ depending on the aggregation level, because households with very different consumptions could end up in the same group. Bolorinos et al. (2020) also found that ML models perform better at a finer spatial scale. They showed that random forests not only outperformed linear models, but also had superior accuracy when predicting water consumption at the individual level. This finding differs from the results of other studies that assessed water consumption at multiple spatial levels (Ouyang et al. 2013). However, this study applied a linear model (linear mixed-effects and ordinary least-squares regression), which has better performance when more spatial homogeneous data are used. For machine learning methods, the amount of data is determinant to model performance, so aggregating information might reduce the learning power of the model. The influence of data set size and the number of variables for ANN models also was pointed out by Lee and Derrible (2020), who showed that fewer explanatory variables are preferred when considering the same data set size.

In terms of R^2, both predictive models were able to explain only part of the residential water demand. Even at the CB scale, at which many variables were available, the best performing model had an R^2 of 0.34. This result suggests that socioeconomic factors alone are not enough to predict water demand, and additional exogenous variables might be necessary. However, there are other possible explanations. The original time series might contain noise or a component that cannot be explained with known variables. Applying a filtering technique before calculating average daily water demand, such as singular spectrum analysis, could solve this problem. In addition, the predictive model could be improved by testing additional statistical learning techniques or by using an ensemble method. Further investigation is recommended to address these issues.

Conclusion

In this study, three ML techniques were used to assess urban water demand in Fortaleza, Brazil: random forests, self-organizing maps, and artificial neural networks. Two spatial levels were addressed: census blocks at the coarse scale, and census tracts at the fine scale.

J. Water Resour. Plann. Manage., 2021, 147(1): 05020026

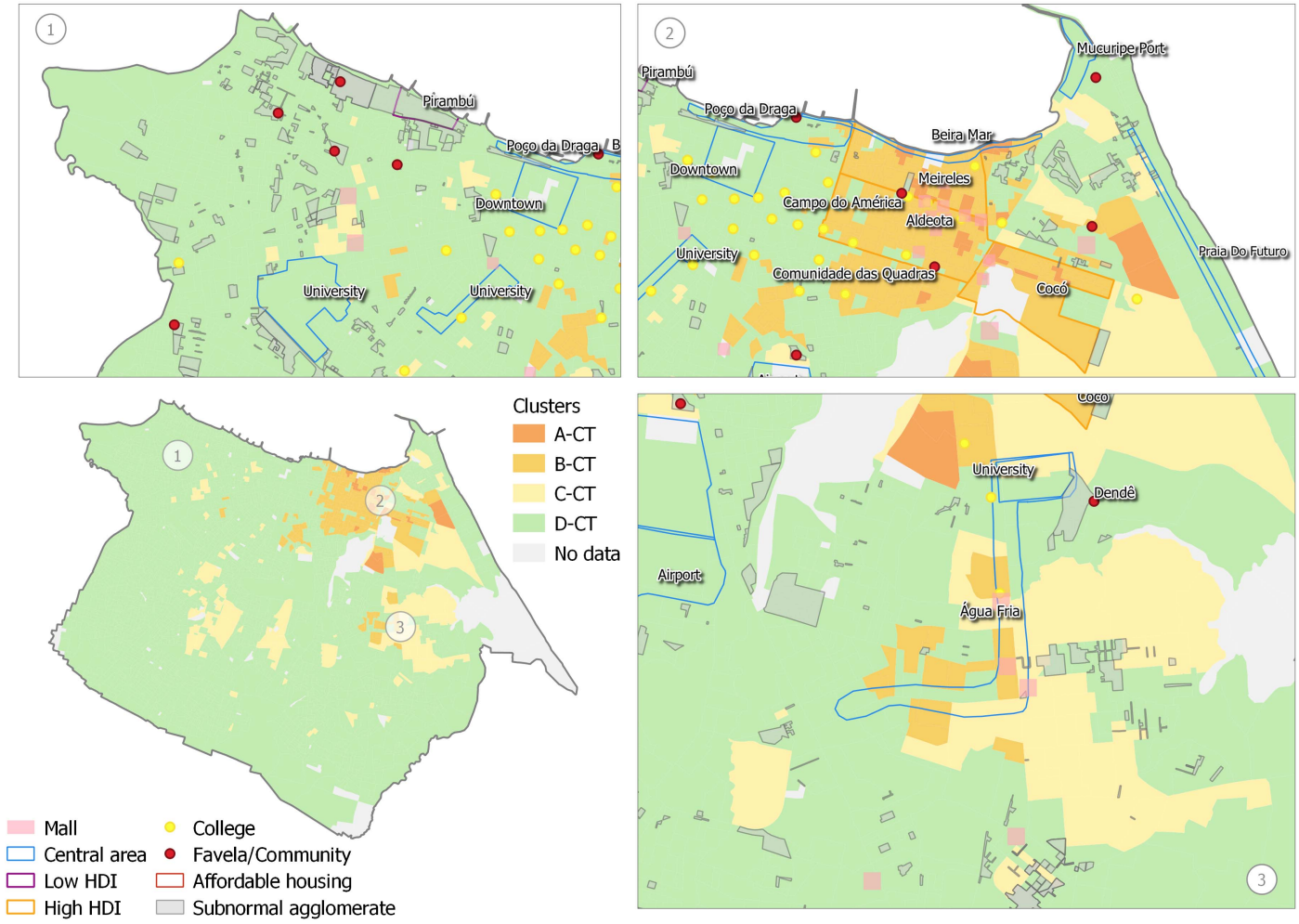

Fig. 11. Clusters defined by the SOM using the explanatory variables of the CT level model (HDI and per capita income).

Table 3. Characteristics of SOM clusters defined using explanatory variables at census tract level

	Cluster			
Variables	A-CT ($n = 24$)	B-CT ($n = 204$)	C-CT ($n = 128$)	D-CT ($n = 2{,}596$)
Total area (km^2)	2,640,250	14,700,981	27,919,873	248,262,919
Total population	16,522	134,297	98,534	2,170,488
Average water demand (L/day)	197.94	182.07	136.80	94.03
MHDI	0.829	0.815	0.362	0.322
APCI (R$)	15,122.85	8,145.50	4,647.49	1,437.09

Note: Except for area and population, variables are represented by mean value for all census tracts in each cluster.

The first had 18 sociodemographic explanatory variables, whereas the second had only 2. A RF model was used to define the most influential variables at the CB level, and this ranking was used for clustering. The IIS method, which was built using a RF and an ANN, was used to choose the best input variables for predicting water demand.

The features with the highest importance included those related to household composition (percentage age 65+, percentage of females, and percentage age 1–14), percent of college graduated inhabitants and life expectancy. The clustering analysis with self-organizing maps provided some interesting insights into the socioeconomic heterogeneity of Fortaleza. There is a distinct spatial gradient across the city in terms of sociodemographic characteristics and water demand: central and eastern zones, with high water demand, have better education and health conditions, whereas southern and western regions, with reduced water demand, have low per capita income and HDI. Nonetheless, heterogeneities in water demand are present in the central areas, and these must be taken into consideration in urban and water resources planning. The input variables selected for the ANN-CB model, with reduced redundancy and maximized information, indicated that average per capita income, percentage age 1–14, and percentage of population vulnerable to poverty provide a fair explanation of water demand in Fortaleza.

The aspects influencing water consumption still are not completely understood, and machine learning (ML) methods are useful

Table 4. Mean of relative error between water demand in census blocks and census tracts and cluster average water demand (%)

Spatial level	Cluster	Mean relative error
Census blocks	A-CB	18.45
	B-CB	20.08
	C-CB	21.05
	D-CB	20.01
	E-CB	17.31
Census tracts	A-CT	53.85
	B-CT	63.38
	C-CT	58.34
	D-CT	43.27

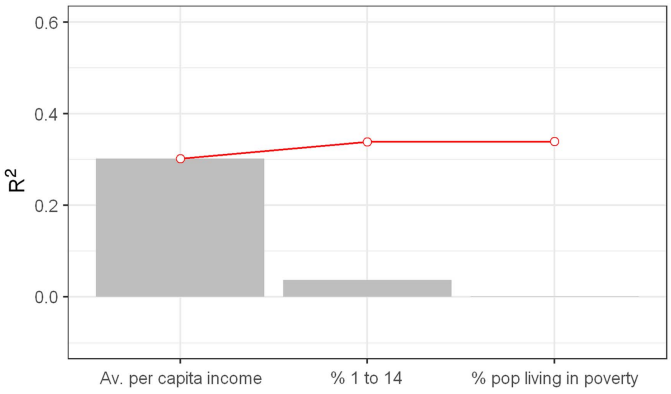

Fig. 12. Increase in performance (R^2) by adding the variables chosen in the iterative input selection method. Bars represent the increase in the R^2 obtained by adding each variable to the input data set, and the line represents the cumulative R^2.

Table 5. ANN-CB (three explanatory variables) and ANN-CT (two explanatory variables) model performance

Performance metric	Aggregation level	
	Census block (CB)	Census tract (CT)
MAE	20.97	22.83
RMSE	31.11	32.38
R^2	0.34	0.43

for identifying behavior patterns. Data availability has a strong influence on the best approach for the modeling. If the data set consists of high-dimensional data (in terms of the number of variables), a variable selection method should be considered. The number of observations can influence model performance; hence, spatially aggregated data might reduce prediction accuracy. However, a coarse scale might provide better insight into spatial analysis of water demand patterns. Features such as the accumulated local effect plots can be useful for interpreting black box models.

This study provided a better understanding of the influence of socioeconomic variables on the water demand of Fortaleza. The results are important not only for prediction, but also for designing targeted water conservation or pricing policies. Further studies could address temporal changes of water demand and scenarios of economic development to support utilities in their long-term planning.

Data Availability Statement

All code used in this study is available online at https://github.com/taiscarvalho/ml_waterdemand. The water demand data were provided by a third party. Direct requests for these data may be made to the Water and Wastewater Company of Ceará (CAGECE).

Acknowledgments

The research was supported by grants from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES) and the Conselho Nacional de Desenvolvimento Científico e Tecnológico—Brasil (CNPq).

Supplemental Materials

Supplemental Texts S1–S4, Figs. S1–S3, and Table S1 are available online in the ASCE Library (www.ascelibrary.org).

References

Adamowski, J., and C. Karapataki. 2010. "Comparison of multivariate regression and artificial neural networks for peak urban water-demand forecasting: Evaluation of different ANN learning algorithms." *J. Hydrol. Eng.* 15 (10): 729–743. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000245.

Altunkaynak, A., and T. A. Nigussie. 2017. "Monthly water consumption prediction using season algorithm and wavelet transform–based models." *J. Water Resour. Plann. Manage.* 143 (6): 04017011. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000761.

Apley, D. W., and J. Zhu. 2016. "Visualizing the effects of predictor variables in black box supervised learning models." Preprint, submitted December 27, 2016. https://arxiv.org/abs/1612.08468.

Barreto, F. A. F. D., and A. S. B. Menezes. 2014. *Desenvolvimento Econômico do Ceará: Evidências Recentes e Reflexões*. Fortaleza, Brazil: Instituto de Pesquisa e Estratégia Econômica do Ceará.

Bata, M. H., R. Carriveau, and S.-K. D. Ting. 2020. "Short-term water demand forecasting using nonlinear autoregressive artificial neural networks." *J. Water Resour. Plann. Manage.* 146 (3): 04020008. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001165.

Bennett, C., R. A. Stewart, and C. D. Beal. 2013. "ANN-based residential water end-use demand forecasting model." *Expert Syst. Appl.* 40 (4): 1014–1023. https://doi.org/10.1016/j.eswa.2012.08.012.

Biau, G., and E. Scornet. 2016. "A random forest guided tour." *Test* 25 (2): 197–227. https://doi.org/10.1007/s11749-016-0481-7.

Bishop, C. M. 1995. *Neural networks for pattern recognition*. Oxford, UK: Oxford University Press.

Bolorinos, J., N. K. Ajami, and R. Rajagopal. 2020. "Consumption change detection for urban planning: Monitoring and segmenting water customers during drought." *Water Resour. Res.* 56 (3): e2019WR025812. https://doi.org/10.1029/2019WR025812.

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

Brentan, B. M., E. Luvizotto Jr., M. Herrera, J. Izquierdo, and R. Pérez-García. 2017. "Hybrid regression model for near real-time urban water demand forecasting." *J. Comput. Appl. Math.* 309 (Jan): 532–541. https://doi.org/10.1016/j.cam.2016.02.009.

Cardell-Oliver, R., J. Wang, and H. Gigney. 2016. "Smart meter analytics to pinpoint opportunities for reducing household water use." *J. Water Resour. Plann. Manage.* 142 (6): 04016007. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000634.

Chang, H., G. H. Parandvash, and V. Shandas. 2010. "Spatial variations of single-family residential water consumption in Portland, Oregon." *Urban Geogr.* 31 (7): 953–972. https://doi.org/10.2747/0272-3638.31.7.953.

Chaudhary, V., R. S. Bhatia, and A. K. Ahlawat. 2014. "A novel Self-Organizing Map (SOM) learning algorithm with nearest and farthest

neurons." *Alexandria Eng. J.* 53 (4): 827–831. https://doi.org/10.1016/j.aej.2014.09.007.

Chen, G., T. Long, J. Xiong, and Y. Bai. 2017. "Multiple random forests modelling for urban water consumption forecasting." *Water Resour. Manage.* 31 (15): 4715–4729. https://doi.org/10.1007/s11269-017-1774-7.

Cominola, A., K. Nguyen, M. Giuliani, R. A. Stewart, H. R. Maier, and A. Castelletti. 2019. "Data mining to uncover heterogeneous water use behaviors from smart meter data." *Water Resour. Res.* 55 (11): 9315–9333. https://doi.org/10.1029/2019WR024897.

Cominola, A., E. S. Spang, M. Giuliani, A. Castelletti, J. R. Lund, and F. J. Loge. 2018. "Segmentation analysis of residential water-electricity demand for customized demand-side management programs." *J. Cleaner Prod.* 172 (Jan): 1607–1619. https://doi.org/10.1016/j.jclepro.2017.10.203.

Dias, T. F., A. Kalbusch, and E. Henning. 2018. "Factors influencing water consumption in buildings in southern Brazil." *J. Cleaner Prod.* 184 (May): 160–167. https://doi.org/10.1016/j.jclepro.2018.02.093.

Duerr, I., H. R. Merrill, C. Wang, R. Bai, M. Boyer, M. D. Dukes, and N. Bliznyuk. 2018. "Forecasting urban household water demand with statistical and machine learning methods using large space-time data: A Comparative study." *Environ. Modell. Software* 102 (Apr): 29–38. https://doi.org/10.1016/j.envsoft.2018.01.002.

Dunn, J. C. 1974. "Well-separated clusters and optimal fuzzy partitions." *J. Cybern.* 4 (1): 95–104. https://doi.org/10.1080/01969727408546059.

Firat, M., M. A. Yurdusev, and M. E. Turan. 2008. "Evaluation of artificial neural network techniques for municipal water consumption modeling." *Water Resour. Manage.* 23 (4): 617–632. https://doi.org/10.1007/s11269-008-9291-3.

Galelli, S., and A. Castelletti. 2013. "Tree-based iterative input variable selection for hydrological modeling." *Water Resour. Res.* 49 (7): 4295–4310. https://doi.org/10.1002/wrcr.20339.

Garcia, J., L. R. Salfer, A. Kalbusch, and E. Henning. 2019. "Identifying the drivers of water consumption in single-family households in Joinville, Southern Brazil." *Water* 11 (10): 1990. https://doi.org/10.3390/w11101990.

Garmany, J. 2011. "Situating Fortaleza: Urban space and uneven development in northeastern Brazil." *Cities* 28 (1): 45–52. https://doi.org/10.1016/j.cities.2010.08.004.

Genuer, R., J.-M. Poggi, and C. Tuleau-Malot. 2010. "Variable selection using random forests." *Pattern Recognit. Lett.* 31 (14): 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014.

Gharabaghi, S., E. Stahl, and H. Bonakdari. 2019. "Integrated nonlinear daily water demand forecast model (case study: City of Guelph, Canada)." *J. Hydrol.* 579 (Sep): 124182. https://doi.org/10.1016/j.jhydrol.2019.124182.

Gonzales, P., and N. Ajami. 2017. "Social and structural patterns of drought-related water conservation and rebound." *Water Resour. Res.* 53 (12): 10619–10634. https://doi.org/10.1002/2017WR021852.

Grande, M. H., C. O. Galvão, L. I. B. Miranda, and L. D. G. Sobrinho. 2016. "The perception of users about the impacts of water rationing on their household routines." *Ambiente Sociedade* 19 (1): 163–182. https://doi.org/10.1590/1809-4422asoc150155r1v1912016.

Gulis, G. 2000. "Life expectancy as an indicator of environmental health." *European J. Epidemiol.* 16 (2): 161–165. https://doi.org/10.1023/A:1007629306606.

Guo, G., S. Liu, Y. Wu, J. Li, R. Zhou, and X. Zhu. 2018. "Short-term water demand forecast based on deep learning method." *J. Water Resour. Plann. Manage.* 144 (12): 04018076. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000992.

Guyon, I., and A. Elisseeff. 2003. "An introduction to variable and feature selection." *J. Mach. Learn. Res.* 3 (Mar): 1157–1182. https://doi.org/10.5555/944919.944968.

Haque, M. M., A. de Souza, and A. Rahman. 2017. "Water demand modelling using independent component regression technique." *Water Resour. Manage.* 31 (1): 299–312. https://doi.org/10.1007/s11269-016-1525-1.

Haselbeck, V., J. Kordilla, F. Krausea, and M. Sauterb. 2019. "Self-organizing maps for the identification of groundwater salinity sources based on hydrochemical data." *J. Hydrol.* 576 (Sep): 610–619. https://doi.org/10.1016/j.jhydrol.2019.06.053.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer Science & Business Media.

Hemati, A., M. A. Rippy, S. B. Grant, K. Davis, and D. Feldman. 2016. "Deconstructing demand: The anthropogenic and climatic drivers of urban water consumption." *Environ. Sci. Technol.* 50 (23): 12557–12566. https://doi.org/10.1021/acs.est.6b02938.

Herrera, M., L. Torgo, J. Izquierdo, and R. Pérez-García. 2010. "Predictive models for forecasting hourly urban water demand." *J. Hydrol.* 387 (1–2): 141–150. https://doi.org/10.1016/j.jhydrol.2010.04.005.

House-Peters, L. A., and H. Chang. 2011. "Urban water demand modeling: Review of concepts, methods, and organizing principles." *Water Resour. Res.* 47 (5): 351–360. https://doi.org/10.1029/2010WR009624.

House-Peters, L. A., B. Pratt, and H. Chang. 2010. "Effects of urban spatial structure, sociodemographics, and climate on residential water consumption in Hillsboro, Oregon." *J. Am. Water Resour. Assoc.* 46 (3): 461–472. https://doi.org/10.1111/j.1752-1688.2009.00415.x.

Hussien, W. A., F. A. Memon, and D. A. Savic. 2016. "Assessing and modelling the influence of household characteristics on per capita water consumption." *Water Resour. Manage.* 30 (9): 2931–2955. https://doi.org/10.1007/s11269-016-1314-x.

IBGE (Instituto Brasileiro de Geografia e Estatística). 2010. "Censo Demográfico de 2010" [2010 Demographic Census]. Accessed November 17, 2018. https://censo2010.ibge.gov.br/.

IPLANFOR (Instituto de Planejamento de Fortaleza). 2015. *Fortaleza 2040*. Fortaleza, Brazil: Prefeitura Municipal de Fortaleza.

Kalteh, A. M., P. Hjorth, and R. Berndtsson. 2008. "Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application." *Environ. Modell. Software* 23 (7): 835–845. https://doi.org/10.1016/j.envsoft.2007.10.001.

Kohonen, T. 1982. "Self-organized formation of topologically correct feature maps." *Biol. Cybern.* 43 (1): 59–69. https://doi.org/10.1007/BF00337288.

Kohonen, T. 2014. *MATLAB implementations and applications of the self-organizing map*. Helsinki, Finland: Unigrafia Oy.

Lee, D., and S. Derrible. 2020. "Predicting residential water demand with machine-based statistical learning." *J. Water Resour. Plann. Manage.* 146 (1): 04019067. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001119.

Li, T., G. Sun, C. Yang, K. Liang, S. Ma, and L. Huang. 2018. "Using self-organizing map for coastal water quality classification: Towards a better understanding of patterns and processes." *Sci. Total Environ.* 628–629 (Jul): 1446–1459. https://doi.org/10.1016/j.scitotenv.2018.02.163.

Lindsay, J., A. J. Dean, and S. Supski. 2017. "Responding to the millennium drought: comparing domestic water cultures in three Australian cities." *Reg. Environ. Change* 17 (2): 565–577. https://doi.org/10.1007/s10113-016-1048-6.

Lippmann, R. 1987. "An introduction to computing with neural nets." *IEEE ASSP Mag.* 4 (2): 4–22. https://doi.org/10.1109/MASSP.1987.1165576.

Liu, Y., J. Zhao, and Z. Wang. 2015. "Identifying determinants of urban water use using data mining approach." *Urban Water J.* 12 (8): 618–630. https://doi.org/10.1080/1573062X.2014.923920.

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.

Matos, C., C. A. Teixeira, R. Bento, J. Varajão, and I. Bentesa. 2014. "An exploratory study on the influence of socio-demographic characteristics on water end uses inside buildings." *Sci. Total Environ.* 466–467 (Jan): 467–474. https://doi.org/10.1016/j.scitotenv.2013.07.036.

Molnar, C. 2019. "Interpretable machine learning." Accessed May 20, 2019. https://christophm.github.io/interpretable-ml-book.

Montgomery, M. A., and M. Elimelech. 2007. "Water and sanitation in developing countries: Including health in the equation." *Environ. Sci. Technol.* 41 (1): 17–24. https://doi.org/10.1021/es072435t.

Msiza, I. S., F. V. Nelwamondo, and T. Marwala. 2007. "Artificial neural networks and support vector machines for water demand time series forecasting." In *Proc., 2007 IEEE Int. Conf. on Systems, Man and Cybernetics*, 638–643. New York: IEEE.

Musolesi, A., and M. Nosvelli. 2007. "Dynamics of residential water consumption in a panel of Italian municipalities." *Appl. Econ. Lett.* 14 (6): 441–444. https://doi.org/10.1080/13504850500425642.

Nawaz, R., P. Rees, S. Clark, G. Mitchell, A. McDonald, M. Kalamandeen, C. Lambert, and R. Henderson. 2019. "Long-term projections of domestic water demand: A case study of London and the Thames Valley." *J. Water Resour. Plann. Manage.* 145 (11): 05019017. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001088.

Olmstead, S. M., and R. N. Stavins. 2009. "Comparing price and nonprice approaches to urban water conservation." *Water Resour. Res.* 45 (4): W04301. https://doi.org/10.1029/2008WR007227.

Ouyang, Y., E. A. Wentz, B. L. Ruddell, and S. L. Harlan. 2013. "A multiscale analysis of single-family residential water use in the Phoenix metropolitan area." *J. Am. Water Resour. Assoc.* 50 (2): 448–467. https://doi.org/10.1111/jawr.12133.

Padulano, R., and G. Giudice. 2018. "A mixed strategy based on self-organizing map for water demand pattern profiling of large-size smart water grid data." *Water Resour. Manage.* 32 (11): 3671–3685. https://doi.org/10.1007/s11269-018-2012-7.

Papacharalampous, G. A., and H. Tyralis. 2018. "Evaluation of random forests and prophet for daily streamflow forecasting." *Adv. Geosci.* 45: 201–208. https://doi.org/10.5194/adgeo-45-201-2018.

PNUD (Programa das Nações Unidas para o Desenvolvimento), IPEA (Instituto de Pesquisa Econômica Aplicada), and FJP (Fundação João Pinheiro). 2014. "Atlas do desenvolvimento humano nas regiões metropolitanas" [Atlas of human development in metropolitan regions]. Accessed November 18, 2018. https://atlasbrasil.org.br.

Polebitski, A. S., and R. N. Palmer. 2010. "Seasonal residential water demand forecasting for census tracts." *J. Water Resour. Plann. Manage.* 136 (1): 27–36. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000003.

Pulido-Calvo, I., P. Montesinos, J. Roldán, and F. Ruiz-Navarro. 2007. "Linear regressions and neural approaches to water demand forecasting in irrigation districts with telemetry systems." *Biosyst. Eng.* 97 (2): 283–293. https://doi.org/10.1016/j.biosystemseng.2007.03.003.

Qi, J., H. Liu, X. Liu, and Y. Zhang. 2019. "Spatiotemporal evolution analysis of time-series land use change using self-organizing map to examine the zoning and scale effects." *Comput. Environ. Urban Syst.* 76 (Jul): 11–23. https://doi.org/10.1016/j.compenvurbsys.2019.03.002.

Quesnel, K. J., and N. K. Ajami. 2017. "Changes in water consumption linked to heavy news media coverage of extreme climatic events." *Sci. Adv.* 3 (10): e1700784. https://doi.org/10.1126/sciadv.1700784.

Rasifaghihi, N., S. S. Li, and F. Haghighat. 2020. "Forecast of urban water consumption under the impact of climate change." *Sustainable Cities Soc.* 52 (Jan): 101848. https://doi.org/10.1016/j.scs.2019.101848.

Reed, R., and R. J. Marksii. 1999. *Neural smithing: Supervised learning in feedforward artificial neural networks*. Cambridge, MA: Massachusetts Institute of Technology Press.

Romano, G., N. Salvati, and A. Guerrini. 2014. "Estimating the determinants of residential water demand in Italy." *Water* 6 (10): 2929–2945. https://doi.org/10.3390/w6102929.

Rousseeuw, P. J. 1987. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis." *J. Comput. Appl. Math.* 20 (Nov): 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams. 1986. "Learning representations by back-propagating errors." *Nature* 323 (6088): 533–536. https://doi.org/10.1038/323533a0.

Sant'Ana, D., and P. Mazzega. 2018. "Socioeconomic analysis of domestic water end-use consumption in the Federal District, Brazil." *Sustainable Water Resour. Manage.* 4 (4): 921–936. https://doi.org/10.1007/s40899-017-0186-4.

Sauri, D. Forthcoming. "The decline of water consumption in Spanish cities: Structural and contingent factors." *Int. J. Water Resour. Dev.* https://doi.org/10.1080/07900627.2019.1634999.

Schleich, J., and T. Hillenbrand. 2009. "Determinants of residential water demand in Germany." *Ecol. Econ.* 68 (6): 1756–1769. https://doi.org/10.1016/j.ecolecon.2008.11.012.

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.

Solomatine, D., L. M. See, and R. J. Abrahart. 2009. "Data-driven modelling: Concepts, approaches and experiences." In *Practical hydroinformatics*, 17–30. Berlin: Springer.

Souza, S., and F. de C. Neves. 2002. *Seca [Drought]*. Fortaleza, Brazil: Edições Demócrito Rocha.

Tyralis, H., G. Papacharalampous, and A. Langousis. 2019. "A brief review of random forests for water scientists and practitioners and their recent history in water resources." *Water* 11 (5): 910–947. https://doi.org/10.3390/w11050910.

UNESCO. 2018. *Nature-based solutions for water: Development report*. Paris: UNESCO.

UNESCO-WWAP (World Water Assessment Programme). 2019. *The United Nations world development report 2019: Leaving no one behind*. Executive Summary. Paris: UNESCO.

Vesanto, J., and E. Alhoniemi. 2000. "Clustering of the self-organizing map." *IEEE Trans. Neural Networks* 11 (3): 586–600. https://doi.org/10.1109/72.846731.

Vijai, P., and P. B. Sivakumar. 2018. "Performance comparison of techniques for water demand forecasting." *Procedia Comput. Sci.* 143: 258–266. https://doi.org/10.1016/j.procs.2018.10.394.

Villarin, M. C., and V. F. Rodriguez-Galiano. 2019. "Machine learning for modeling water demand." *J. Water Resour. Plann. Manage.* 145 (5): 04019017. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001067.

Witten, I. H., and E. Frank. 2016. *Data mining: Practical machine learning tools and techniques*. Burlington, MA: Morgan Kaufmann.

Xiao, Y., L. Fang, and K. W. Hipel. 2018. "Agent-based modeling approach to investigating the impact of water demand management." *J. Water Resour. Plann. Manage.* 144 (3): 04018006. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000907.

Yajima, H., and J. Derot. 2017. "Application of the Random Forest model for chlorophyll-*a* forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases." *J. Hydroinf.* 20 (1): 206–220. https://doi.org/10.2166/hydro.2017.010.

Yousefi, P., S. Shabani, H. Mohammadi, and G. Naser. 2017. "Gene expression programing in long term water demand forecasts using wavelet decomposition." *Procedia Eng.* 186: 544–550. https://doi.org/10.1016/j.proeng.2017.03.268.

Ziegler, A., and I. R. König. 2014. "Mining data with random forests: current options for real-world applications." *Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery* 4 (1): 55–63. https://doi.org/10.1002/widm.1114.