UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA HIDRAÚLICA E AMBIENTAL
PROGRAMA DE PÓS-GRADUAÇÃO EM RECURSOS HÍDRICOS

TAÍS MARIA NUNES CARVALHO

WATER DEMAND MODELLING USING MACHINE LEARNING TECHNIQUES

FORTALEZA
2019

TAÍS MARIA NUNES CARVALHO

WATER DEMAND MODELLING USING MACHINE LEARNING TECHNIQUES

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Civil da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Mestra em Engenharia Civil. Área de concentração: Recursos Hídricos.

Orientador: Prof. Dr. Francisco de Assis de Souza Filho.

FORTALEZA

2019

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

C329w Carvalho, Taís Maria.
Water demand modeling using machine learning techniques / Taís Maria Carvalho. – 2019.
60 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia Civil: Recursos Hídricos, Fortaleza, 2019.
Orientação: Prof. Dr. Francisco de Assis de Souza Filho.

1. Demanda de água. 2. Seleção de variáveis. 3. Aprendizado de Máquina. 4. Previsão de longo prazo. I. Título.

CDD 627

TAÍS MARIA NUNES CARVALHO

WATER DEMAND MODELLING USING MACHINE LEARNING TECHNIQUES

Dissertation presented to the Graduate Program in Civil Engineering of the Federal University of Ceará, as a partial requirement to obtain a Master's degree in Civil Engineering. Concentration area: Water resources.

Approved in: 11/03/2019.

EXAMINATION BOARD

____
Prof. Dr. Francisco de Assis de Souza Filho (Advisor)
Universidade Federal do Ceará (UFC)

____
Prof^{a}. Dr^{a}. Iana Alexandra Alves Rufino
Universidade Federal de Campina Grande (UFCG)

____
Prof. Dr. Carlos de Oliveira Galvão
Universidade Federal de Campina Grande (UFCG)

AGRADECIMENTOS

Aos meus pais, Terezinha e Cláudio, que nunca mediram esforços para que eu pudesse ter acesso a uma educação de qualidade. Agradeço pelo carinho, cuidado e atenção que foram fundamentais durante a minha vida acadêmica. Eu não teria chegado tão longe sem esse suporte e incentivo.

Aos meus irmãos, Taiana e João, pela parceria e presença em todos os momentos. Agradeço por todo o cuidado que tem comigo e por serem meus melhores amigos.

Ao meu companheiro, Walter, pelo constante incentivo e paciência durante todo o processo. Agradeço pelo carinho, parceria e por ser sempre inspiração para mim.

Ao meu orientador e amigo, Prof. Francisco de Assis de Souza Filho, pelo constante incentivo e por ser inspiração profissional e pessoal.

Aos participantes da banca examinadora Prof. Iana Alexandra Alves Rufino e Prof. Carlos de Oliveira Galvão, agradeço pela disponibilidade para avaliar e contribuir com este trabalho.

Aos amigos do grupo de pesquisa Gerenciamento do Risco Climático (GRC) pelo companheirismo, apoio e parceria, em especial Gabriela, Victor, Renan, Larissa e Renata.

Aos amigos que sempre estiveram presentes e que tornaram a caminhada mais leve, em especial Cida, Priscila, Renata e Suyanne.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo financiamento da bolsa de estudos durante o curso de Mestrado.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

RESUMO

A previsão da demanda de água é fundamental para decisões relacionadas à gestão de recursos hídricos a longo prazo. No entanto, a variabilidade espacial do consumo de água é um desafio para uma previsão adequada. O principal objetivo do presente estudo é investigar como os aspectos socioeconômicos da população afetam o futuro consumo urbano de água. Para que a previsão tenha bom desempenho e precisão, um subconjunto significativo de variáveis explicativas deve ser definido. Para isso, vários métodos de seleção de variáveis do tipo filtro e envoltório, baseados Regressão de Mínimos Quadrados Parciais (PLSR, do inglês Partial Least Square Regression) foram testados, juntamente com uma classificação baseada em Florestas Aleatórias (RF, do inglês Random Forest). Os subconjuntos de dados foram em seguida utilizados como entrada para um modelo preditivo. Duas técnicas de aprendizado de máquina foram testadas: RF e Rede Neural Artificial (RNA). O desempenho do modelo foi avaliado através do coeficiente de Nash-Sutfcliffe, Raiz do erro quadrático médio (RMSE, do inglês Root Mean Square Error) e correlação de Pearson. O conjunto de dados consistiu no consumo de água e dados do Censo de 2010 associados a 182 Unidades de Desenvolvimento Humano (UDH) em Fortaleza, Ceará. Importância da variável em projeção, Procedimento de eliminação regularizada e RF forneceram os subconjuntos de variáveis que levaram ao melhor desempenho de previsão entre os sete métodos de seleção. A expectativa de vida ao nascer, a renda per capita e residentes com educação primária e secundária foram considerados variáveis importantes na maioria dos subgrupos. De acordo com a avaliação de desempenho, os modelos RNA e PLSR tiveram melhor desempenho do que a RF na previsão da demanda de água. O RMSE para o melhor modelo PLSR foi de 25.779 Litros/pessoa/dia (Lpd^{-1}) e 24.776 Lpd^{-1} para RNA, enquanto para RF 31.820 Lpd^{-1}. Variáveis socioeconômicas apresentaram grande influência no consumo de água, com destaque para renda per capita e escolaridade. Embora frequentemente usada para previsão de curto prazo, a RNA mostrou-se uma boa abordagem para a previsão da demanda de água a longo prazo. A abordagem proposta para projetar um modelo espacial de consumo de água pode ser estendida a outras regiões metropolitanas e diferentes conjuntos de dados.

Palavras-chave: Demanda de água. Seleção de variáveis. Aprendizado de Máquina. Previsão de longo prazo.

ABSTRACT

Water demand forecasting is fundamental to decisions related to long-term water resources management. However, spatial variability of water consumption may turn prediction into a difficult task. The main purpose of the current study is to investigate how socioeconomic aspects of households affect future urban water consumption. Prior to designing the prediction model, a significant subset of explanatory variables had to be chosen for an improved performance and accuracy. Therefore, several filter and wrapper variable selection methods in Partial Least Squares Regression (PLSR) were tested, along with a classification based on Random Forests (RF). The feature subsets were used as input for a predictive model. Two machine learning techniques were tested: RF and Artificial Neural Network (ANN). Model performance was evaluated through Nash-Sutfcliffe coefficient, Root Mean Square Error (RMSE) and Pearson Correlation. The dataset consisted in 2010 water consumption and Census data associated with 182 Human Development Units (HDU) in Fortaleza, Ceará. Variable importance in projection (VIP), Regularized elimination procedure (REP-PLS) and RF provided the variable subsets that led to the best prediction performance among the seven selection methods. Life expectancy at birth, per capita income and residents with primary and secondary education were considered as important variables in most of the feature subsets. According to the performance assessment, ANN an PLSR provided similar performances and better estimates than RF in predicting water demand. RMSE for the best PLSR model was 25.779 Liters/person/day (Lpd^{-1}) and 24.776 Lpd^{-1} for ANN, while for RF 31.820 Lpd^{-1}. Socioeconomic variables presented great influence in water consumption, especially per capita income and education. Although frequently used for short-term forecasting, ANN was proved a good approach for long-term water demand prediction. The proposed approach of designing a spatial water consumption model can be extended to other metropolitan regions and different datasets.

Keywords: Water demand. Variable selection. Machine Learning. Long-term Prediction.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ANN	Artificial Neural Network
BVE-PLS	Backward variable elimination
CAGECE	Water and Wastewater Company of Ceará
CART	Classification and Regression tree Models
HDU	Human Development Unit
IBGE	Brazilian Institute of Geography and Statistics
IPEA	Institute of Applied Economic Research
MHDI	Municipal Human Development Index
MRF	Metropolitan Region of Fortaleza
NSE	Nash-Sutcliffe efficiency
NRMSE	Normalized Root Mean Square Error
OOB	Out-of-bag
PCC	Pearson correlation coefficient
PLS	Partial Least Squares
PLSR	Partial Least Squares Regression
REP-PLS	Regularized elimination procedure
RF	Random Forest
RMSE	Root Mean Square Error
RMSEP	Root Mean Squared Error of Prediction
SR	Selectivity ratio
SwPA-PLS	Sub-window permutation analysis coupled with PLS
UVE-PLS	Uninformative variable elimination
VIP	Variable importance in projection

SUMMARY

1 INTRODUCTION

The management of water resources systems involves an adequate supply and demand assessment. Water availability uncertainties associated with climate and socio-economic changes imply an increase of supply costs. Therefore, accurate forecasting of short and long-term water demand is fundamental for strategic planning and operation of water systems.

Fortaleza has faced a growing urban water demand and recurrent droughts during the past years. The projected population and economic growth of the city requires the expansion of water supply sources and distribution facilities. In order to develop an expansion strategy and properly allocate resources, managers need disaggregated information on water consumption.

While short-term water forecasting is useful for operational decisions, long-term forecasting is mandatory for planning and design of water supply (HERRERA et al., 2010; BOUGADIS et al., 2005). Although there is not consensus about the time frame for these horizons, usually predicting demand in a long-term horizon means forecasts for 10 years or more, while hourly to monthly forecasts are classified as short-term predictions (GARDINER; HERRINGTON, 2014).

Water demand might be affected by different variables depending on the planning horizon. Socioeconomic changes of population affect water demand slowly over a period of years. Climate factors produce a seasonal influence on demand, while rainfall, temperature fluctuations and stochastic events produce immediate fluctuations in demand (MAIDMENT; MIAOU, 1986; ZHOU et al., 2000).

Water demand is usually predicted for an aggregate level, considering the water use within a city is uniform. However, researchers have investigated water use at the household (WENTZ; GOBER, 2007; DUERR et al., 2018) and census tract level (LEE; WENTZ, 2008; POLEBITSKI; PALMER, 2010). These predictions provide more information about the consumers and the identification of spatial patterns of demand.

The decision about which method to use for urban water demand forecasting depends on the data available and the planning horizon. Some of the techniques commonly used for long-term prediction are regression (BREKKE et al., 2002; POLEBITSKI; PALMER, 2010), scenario-based (GOODCHILD, 2003) and time-series analysis (ALHUMOUD, 2008).

Bougadis et al. (2005) explored regression, time series, and ANN models for weekly water demand. Wang et al. (2009) combined ANN and regression for long-term prediction. Duerr et al. (2018) compared three machine learning techniques for water demand forecasting to an autoregressive model (AR): Random Forest (RF), Bayesian additive regression trees (BART) and gradient boosting algorithms (GBM). The AR model provided the most accurate forecast, although RF provided the best long-term uncertainty quantification.

Although Artificial Neural Networks (ANN) are mainly applied to short term demand forecast (BOUGADIS et al., 2005; ADAMOWSKI, 2008; BENNETT et al., 2013), they can be used to determine the relationship between dependent and independent variables (BEHBOUDIAN et al., 2013).

In this research, Multilayer Perceptron Artificial Neural Networks and Random Forest predictive models were developed and compared with a conventional linear method, the Partial Least Squares Regression (PLSR). A cross-section analysis was performed, where the socioeconomic characteristics of different spatial units are used to estimate water demand. In addition, different methods of variable selection were applied to choose the subset of explanatory variables.

2 OBJECTIVES

2.1 Main Goal

Develop a spatial prediction model for urban water demand using socioeconomic data for a long-term planning horizon applying machine learning techniques.

2.2 Specific Goals

- Analyze water demand patterns at the spatial scale of Human Development Units and quantify the relationship between socioeconomic variables and urban water consumption.
- Classify the socioeconomic variables according to their importance level for forecasting water demand.
- Identify the best socioeconomic variable subsets for long-term water demand forecasting.
- Develop a model for water demand forecasting at the spatial scale of Human Development Units.

3 METHODS

In this section, the study area, data, variable selection methods and predictive models used in this research are presented.

3.1 Study Area

Fortaleza, capital of Ceará, is a semi-arid region of Northeastern Brazil and is part of the Metropolitan Region of Fortaleza (MRF). The MRF water supply system consists of eight storage reservoirs, pump stations and canals that transfer water from the Jaguaribe River basin, where is located the largest reservoir of the state (6.2 billion m^3). Five reservoirs are in the MRF (Gavião, Pacoti-Riachão, Pacajus and Aracoiaba) and three in the Jaguaribe Basin (Orós, Castanhão and Banabuiú). In the future, the system will also be supplied by São Francisco river transposition (PREFEITURA MUNICIPAL DE FORTALEZA, 2016).

The total capacity of the reservoirs is 8,002 hm^3 and the water consumption is estimated at 45.30 m^3/s, with Jaguaribe region accounting for 71% of total demand. Metropolitan basin demand corresponds to domestic, municipal and industrial uses. West of MRF, the Industrial and Port Complex of Pecém (CIPP) is considered the main industrial consumer, with a water consumption of 1.4m^3/s (PREFEITURA MUNICIPAL DE FORTALEZA, 2016).

Fortaleza's residential average per capita consumption was 116.24 Liters/person/day (Lpd^{-1}) over the 2009-2017 period, reaching its peak of 129.05 Lpd^{-1} in 2013. Population is expected to grow from 2.5 million people in 2012 to 3.14 in 2040, due to a positive net migration rate and a reduction of child mortality and death from external causes (PREFEITURA MUNICIPAL DE FORTALEZA, 2016). Industrial, commercial and agricultural growth will further increase the demand for water.

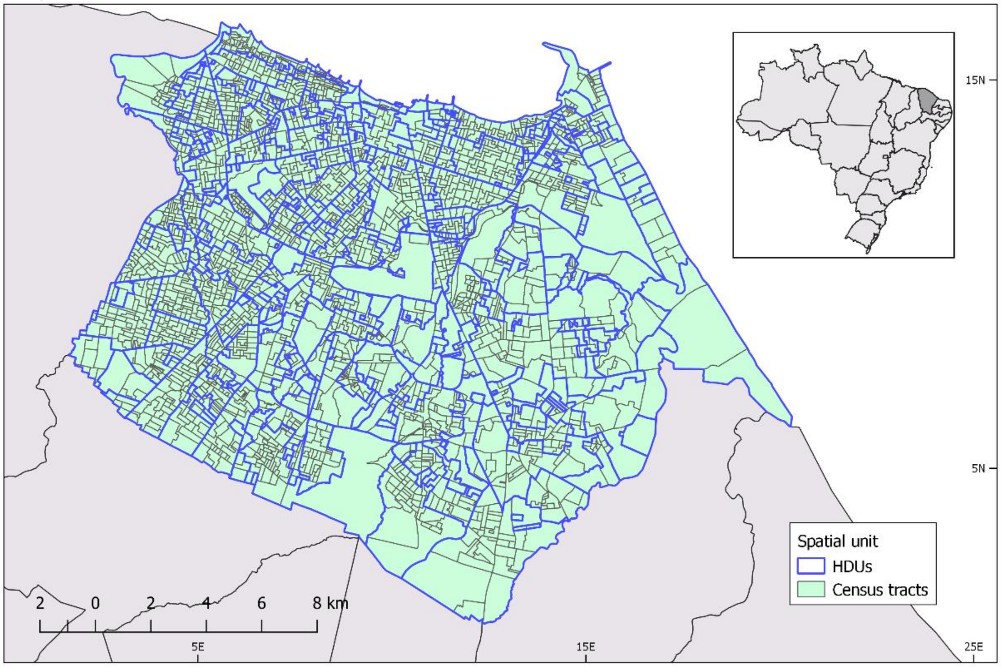
3.2 Data

Average daily per capita water consumption was derived from water consumption data provided by Water and Wastewater Company of Ceará (CAGECE). CAGECE provided commercial, industrial, public and residential monthly water consumption with a household identifier for the period of January 2009 to December 2017. To create daily per capita demand, residential water consumption was aggregated by census tract and divided by its population and number of days for each month.

Socioeconomic and population data used in this research was obtained from the 2010 Census, conducted by the Brazilian Institute of Geography and Statistics (IBGE). IBGE's census data is collected at the level of households and people, but it is only released under rigorous criteria of aggregation and statistical reliability, to avoid the exposure of custom information. Partners of Human Development Atlas in Brazil proposed a spatial configuration called Human Development Units (HDU), that aggregates Census data of the main Brazilian metropolitan areas (PNUD; IPEA; FJP; 2014).

Each HDU has a minimum of 400 permanent and private residences and are as homogenous as possible. They are also recognized by the resident population. These units were validated by local partners with the support of the Institute of Applied Economic Research (PNUD; IPEA; FJP; 2014). Figure 1 represents the intersection between census tracts (n = 2952) and HDUs (n = 182).

Figure 1 – HDUs and census tracts.



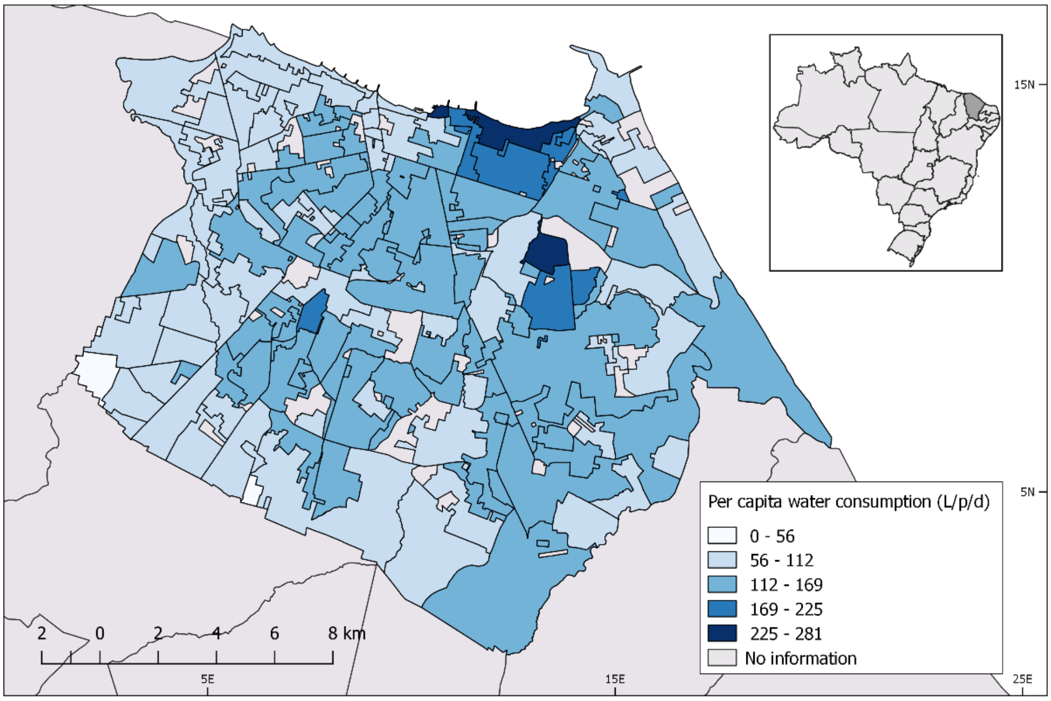
Source: Elaborated by the author.

Since socioeconomic data was obtained at HDU level, water consumption, initially distributed by census tract, was spatially aggregated into HDUs by identifying the spatial intersection between these two units. Then, water demand was weighted by the population of the census tracts within each HDU, as presented by Equation 1.

$$D_i = \frac{\sum_{j=1}^{n} D_j P_j}{\sum_{j=1}^{n} P_j} \qquad (1)$$

Where D_j and P_j are the per capita water demand and population of a census tract *j* and *n* is the number of census tracts contained in an HDU *i*. Data from 182 HDU units was obtained. Figure 2 presents the per capita water consumption in Liters/person/day for each HDU of Fortaleza.

Figure 2 – Per capita water usage in Fortaleza in 2010.



Source: Elaborated by the author.

3.3 Variable Selection

The use of household-level data to estimate residential water demand provides a better understanding of consumption spatial patterns and is preferred rather than aggregate city-scale information. Previous studies that used this approach tried to relate water demand with structural, social and environmental variables, such as lot size, building density, water price, temperature, educational level and family size (CHANG; PARANDVASH; SHANDAS, 2010; ARBUES; VILLANÚA, 2006; POLEBITSKI; PALMER, 2010).

Aggregating data into HDU units was an advantage because (1) they are a better representation of the intra-metropolitan inequalities than IBGE’s census tracts, (2) a large set of socioeconomic information was available, allowing several variables to be tested and classified according to their importance, (3) demand heterogeneity across the city could addressed in the forecast and (4) the model can be reproduced and tested for other metropolitan regions comprised by Atlas Brazil.

Since climatic variables are more likely to induce short-term seasonal variations in water demand (MIAOU, 1990), they were not considered in this model. The explanatory variables (Table 1) were chosen for assessing the role of socioeconomic variables in long-term demand forecasting.

Table 1 – Explanatory variables.

Variable ID	Description	Unit
V01	Per capita income	R$
V02	Life expectancy at birth	Years
V03	Employment rate - 18 years old and older	%
V04	Gini Index	N/A
V05	% of people in households with bathrooms and running water	%
V06	% 5 to 6 years old enrolled in school	%
V07	% 6 to 14 years old enrolled in school	%
V08	% 11 to 13 years old enrolled in final years or that finished Elementary School	%
V09	% 18 to 20 years old that finished High School	%
V10	% 15 to 17 years old that finished Elementary School	%
V11	% 18 or older that finished Elementary School	%
V12	Expected years of schooling	Years
V13	% 1 to 14 years old	%
V14	% 65 years old or above	%
V15	% Male residents	%
V16	% Female residents	%

Source: Elaborated by the author.

Following the criteria for the calculation of the Municipal Human Development Index, indicators related to its three dimensions (longevity, education and income) were included as independent variables of the model. In addition, Gini index, employment rate and percentage of people in households with bathrooms and running water were added to reflect economic aspects of households in each HDU. Variables related to population age and sex were also included.

Prior to the prediction model development, the number of explanatory variables had to be reduced, therefore future water demand scenarios could be more accurate and easier to estimate. There is an extensive discussion in the literature regarding whether to choose feature subsets based on their relevance or usefulness (JOHN; KOHAVI; PFLEGER, 1994; KOHAVI; JOHN, 1997; BLUM; LANGLEY, 1997; LIU; YU, 2005).

Ranking the most relevant variables could result in a subset with redundant features, since selection criteria is based only on their association to the predictor. However, this approach is useful for interpretation and filtering the least promising variables (GUYON;

ELISSEEFF, 2003). Alternatively, the problem could be switched to finding an optimal subset of features that together have a good predictive power (KOHAVI; JOHN, 1997).

John, Kohavi and Pfleger (1994) divide variable selection methods into three classes: wrappers, filters, and embedded. Wrappers classify variable subsets according to their prediction performance, while filters do not consider any model outputs and embedded include variable selection during model estimation.

Wrappers often give better results (in terms of the final predictive accuracy of a learning algorithm) than filters because feature selection is optimized for a specific learning algorithm. However, since a learning algorithm is called to each set of features, wrappers are slow to execute, especially for large databases. Besides, a subset of useful variables may exclude many redundant, but relevant, variables (GUYON; ELISSEEFF, 2003).

Filters usually provide a ranked list of variables which are highly correlated with the predictors, scoring them individually and independently of each other. The final choice of the features set is left to the user. In some cases, the user must specify how many features are required or set a threshold by which feature selection terminates (HALL, 1999). Despite filters limitation in providing an accurate selection for prediction purposes, they can be used as a preprocessing step for a wrapper, reducing space dimensionality (GUYON; ELISSEEFF, 2003).

In this research, a two-stage feature selection approach was formulated by combining filter and wrapper methods and a more sophisticated non-linear feature selector (Random Forest). The subsets served as input for the selected predictive model to evaluate their performance. The parameters for evaluation were the Pearson correlation coefficient (PCC), Nash-Sutcliffe efficiency (NSE), Root Mean Square Error (RMSE) and Normalized Root Mean Square Error (NRMSE).

Pearson correlation coefficient ranges from +1 to −1 (Equation 2). The closer PCC is to 1, stronger is the positive linear correlation, and the closer to −1, stronger is the negative linear correlation. A PCC of 0 means no correlation.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{2}$$

Where n is the number of observations, xi and yi are individual observations, x and y are the mean of the observations.

The direction and magnitude of correlation between average water demand and each variable was initially investigated through PCC. Besides, hierarchical clustering order of the correlation matrix provided information for an empirical selection of the variable subset for prediction.

Nash–Sutcliffe compares the output values of a predictive model to the mean of observed values (Equation 3). NSE ranges from −∞ to 1, where 1 means the predicted and observed values are the same and 0 indicates the predictive model is as good as using the mean of observed values. When NSE is a negative value, the observed mean is a better predictor than the model.

$$NSE = 1 - \frac{\sum_{t=1}^{T}\left(Q_m^t - Q_o^t\right)^2}{\sum_{t=1}^{T}\left(Q_o^t - \overline{Q_o}\right)^2} \tag{3}$$

Where Q$_m$ is the value predicted by the model, Q$_{ot}$ is the observed value at time t and Q$_o$ is the mean of observed values. The RMSE corresponds to the square root of the average of squared errors (Equation 4). Its value is always positive, and a value of 0 would indicate a perfect model. The lower the value of RMSE, the better is a model.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y_i} - y_i)^2}{n}} \tag{4}$$

Where $\hat{y_i}$ corresponds to the predicted values, y$_i$ is the observed values and n is the number of observations. The RMSE divided by the mean of observed values (Equation 5) is the normalized RMSE (NRMSE). Lower values of NRMSE indicate less residual variance.

$$NRMSE = \frac{RMSE}{\bar{y}} \tag{5}$$

Because variables do not commensurate, data was normalized by scaling between 0 and 1 (Equation 6) before applying variable selection methods.

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{6}$$

3.3.1 Variable importance in projection (VIP)

Wold, Johansson and Cocchi (1993) proposed a variable importance measure in PLSR projections named Variable Importance in Projection (VIP). VIP summarizes the influence of each predictor variable on the PLSR model. VIP scores are calculated as the weighted sum of squares of the PLSR weights, which consider the amount of variance explained by each PLSR component (FARRÉS et al., 2015). PLSR combined with the VIP scores is often used when multicollinearity is present among variables (CHONG; JUN, 2005). The VIP score is presented by Equation 7.

$$VIP_j = \sqrt{\frac{\sum_{f=1}^{F} w_{jf}^2 * SSY_f * J}{SSY_{total} * F}} \tag{7}$$

Where w_{jf} is the weight value for variable j and component f, SSY_f is the sum of squares of explained variance for the fth component and J the number of X variables. SSY_{total} is the total sum of squares explained of the dependent variable, and F is the total number of components. The VIP gives the importance of the jth variable in each fth component.

Hence, VIP is a measure of the contribution of each variable according to the variance explained by each PLSR component. Since the average of squared VIP scores equals 1, greater than one rule was used as a criterion for variable selection (CHONG; JUN, 2005). However, this is not a statistically justified limit, and it can be proved very sensitive to the presence of non-relevant information (TRAN et al., 2014).

3.3.2 Selectivity ratio (SR)

Target projection (TP) with selectivity ratio (SR) is a popular method for variable selection in multivariate data analysis, especially useful for prediction. TP reveals the y relevant variation in the X variables captured by a multicomponent PLSR model on a single latent variable. The corresponding TP score vector is proportional to the predicted response (KVALHEIM, 2010).

SR of a variable is obtained by calculating the ratio between the explained and the residual (unexplained) variance of the X variables on the y TP component (Equation 8). This TP utilizes both the predictive ability (regression vector) and the explanatory ability (spectral variance/covariance matrix) for the calculation of the SR (FARRÉS et al., 2015).

$$SR_i = \frac{SS_{i,explained}}{SS_{i,residual}} \tag{8}$$

The threshold for selecting the features with high discriminating ability was defined through an F-test (95%) (RAJALAHTI et al., 2009). The calculated F value (F_{calc}), which is equal to SR_i, must exceed the critical value for the F distribution (Equation 9).

$$F_{calc} = SR_i > F_{crit} = F(\alpha, N-2, N-3) \tag{9}$$

Where N is the number of observations and α the significance level, set to 0.05 in this research. The number of components was also set at 5.

3.3.3 Wrapper PLSR-based methods

Four PLSR-based wrapper methods were tested: Backward variable elimination, Uninformative variable elimination, Sub-window permutation analysis and Regularized elimination procedure.

3.3.3.1 Backward variable elimination (BVE-PLS)

Backward variable elimination (BVE) results in the elimination of non-informative variables (FRANK, 1987). Variables were first sorted according to VIP importance measure. Secondly, a threshold was used to eliminate a subset of the least informative variables (greater than one rule was maintained). Then, PLSR is fitted again to the remaining variables and performance is measured. 75% of the samples was used for calibration. The procedure is repeated until maximum model performance is achieved (MEHMOOD et al., 2012).

3.3.3.2 Uninformative variable elimination (UVE-PLS)

Uninformative variable elimination in PLS (UVE-PLS) consists in adding artificial noise variables that have the same variability of the original ones before the PLSR model is fitted (CENTNER et al., 1996). The X variables of lower importance than the artificial noise variables are eliminated before the procedure is repeated until the performance of the models start decreasing. The elimination of uninformative variables can improve predictive ability, since they do not contain more information than random variables (CENTNER et al., 1996).

The threshold was set as the maximum of absolute value c (leave-one-out cross-validation result) among the noise variables. Again, 75% of the samples were used for calibration and three Monte Carlo sampling simulations were performed.

3.3.3.3 Sub-window permutation analysis coupled with PLS (SwPA-PLS)

Sub-window permutation analysis coupled with PLSR (SwPA-PLS) provides the influence of each variable without considering the influence of the other variables (MEHMOOD et al., 2012). SwPA is suitable for analyzing both datasets with many variables and small ones. In this method, a conditional probability value (P value) is computed for each predictor using Mann–Whitney U test. The P value can be subsequently utilized as a criterion to assess the importance of each variable (LI et al., 2010). 80% of the samples were used for calibration and 3 Monte Carlo sampling simulations were performed. The threshold for variable selection was defined as P = 0.05.

3.3.3.4 Regularized elimination procedure (REP-PLS)

This method performs a parsimonious selection achieved by tolerating a minor performance deviation from any optimum if it results in a reduced number of selected variables (MEHMOOD et al., 2011). A stability-based selection procedure is adopted, where the samples are split randomly into a predefined number of training and test sets. For each split, a stepwise procedure is carried out. 75% of the samples were used for calibration and VIP threshold was set as 1. The threshold for variable selection was defined as P = 0.05.

The PLSR-based methods were implemented using R plsVarSel package (MEHMOOD et al., 2012). These techniques are widely applied in chemometrics and bioinformatics fields and recently have proven useful for a variety of data types. More detailed information about them can be found at Mehmood et al. (2012).

3.3.4 Random Forest

Feature selection using Random Forests (RF) is an embedded method, performing variable selection as part of its learning process. RF is a popular machine learning algorithm for variable selection introduced by Breiman (2001). This method is based on the combination of many classification and regression tree models (CART) trained with bootstrapping aggregation (bagging). The combined result of many decision trees (forest) is used for prediction. For a detailed description of the algorithm, see Breiman (2001).

The good performance of RF is related to its randomness when creating the trees. At each node of the decision tree, the model chooses the best split from a random subset of explanatory variables. In addition, when bootstrapping the training set for each tree, about one third of the observations are left out for performance evaluation, the out-of-bag sample (OOB). The OOB sample is used to get an unbiased estimate of the prediction error (GENUER; POGGI; TULEAU-MALOT, 2010).

To assess the importance of a specific predictor variable, the values of the variable are randomly permuted for the out-of-bag observations, and then the modified out-of-bag data are passed down the tree to get new predictions. Therefore, if a predictor is important for the model, randomly assigning other values for that variable should have a negative influence on prediction.

The difference between the mean of the error (misclassification rate for classification and mean squared error (MSE) for regression) for the modified and original out-of-bag data, divided by the standard error, is a measure of the importance of the variable (CUTLER, 2007).

Another measure is the increase in Node Purity, that relates to the loss function which by best splits are chosen. For classification, the loss function is the Gini impurity, while for regression is variance. The more useful is the variable, higher is the increase in node purity.

The increase in MSE (IncMSE) is a more robust and informative measure than increase in Node Purity. Higher values of IncMSE indicate variable is important for regression. The mean of squared residuals is given by Equation 10.

$$MSE_{OOB} = n^{-1} \sum_{1}^{n} \{y_i - \hat{y}_i^{OOB}\}^2 \qquad (10)$$

Where $\hat{y}_i^{OOB}$is the average of the OOB predictions for the *i*th observation (LIAW; WIENER, 2002). IncMSE was taken as the measure for variable importance representing RF classification.

The method was implemented using the R randomForest package (LIAW; WIENER, 2002) which is based on Breiman's classic algorithm. The two main parameters are “mtry”, the number of input variables randomly chosen at each split and “ntree”, the number of trees in the forest. The assumed value of “mtry” was the default for regression models, the number of explanatory variables divided by 3 (rounded down). The RF model was used for two applications: variable selection and water demand prediction.

3.5 Predictive models

Besides Random Forest, two predictive models were evaluated: Partial Least Squares Regression and Artificial Neural Network. The best model was used for comparing feature subsets.

3.5.1 Partial Least Squares Regression (PLSR)

Partial least squares regression (PLSR) is a multivariate linear regression technique used to find the correlation between a matrix X of predictor variables and a matrix or a vector y of response variables. PLSR extracts linear combinations of the predictors (components), providing information about the correlation structure and behavior of X and y. This method is useful to analyze strongly correlated data, noisy, and numerous explanatory variables (WOLD; SJÖSTRÖM; ERIKSSON, 2001).

The construction of components is the major point of PLSR. The components are the linear transformations of X which maximize covariance between response variable y and components. The approach of finding each component is done sequentially. The first component ($t_1 = Xw_1$) is determined by maximizing the covariance between y and t_1 under the constraint of $||w_1|| = 1$. To extract the other components, original matrix X and y must be reconstructed by substituting of their residuals. This process is called deflation of matrices X and y (AKARACHANTACHOTE; CHADCHAM; SAITHANU, 2014). The residuals of X and y for the first component are found out as of Equation 11 and Equation 12, respectively.

$$\mathrm{E_1 = X - t_1 p_1'} \tag{11}$$

$$\mathrm{f_1 = y - t_1 q_1} \tag{12}$$

Where p_1 and q_1 are loadings defined by OLS fitting. Also, the residual of *a*th components X and y are computed as of Equation 13 and Equation 14, respectively.

$$\mathrm{E_a = E_{a-1} - t_a p_a'} \tag{13}$$

Where $E_0 = X$ and $f_0 = y$.

$$f_a = f_{a-1} - t_a q_a \quad (14)$$

There are various approaches of PLSR, and more detailed variants of PLS can be found in Rosipal and Krämer (2005). In this research, the number of components was first determined as six.

3.5.2 Artificial Neural Network (ANN)

Many methods are used for modeling urban water demand, such as multiple regression (GATO et al., 2007; MAIDMENT, MIAOU, 1986), dynamic models (ROSENBERG, 2007), ARIMA (BOUGADIS et al. (2005) and artificial neural networks (ANN) (GHIASSI; ZIMBRA; SAIDANE, 2008; LI; HUICHENG, 2010).

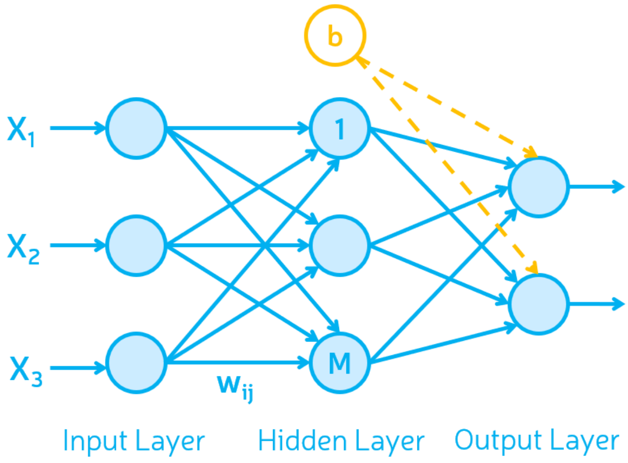
ANN is a powerful technique for non-linear models (ÖZESMI; ÖZESMI, 1999) and a useful tool for forecasting demand in complicated water systems (HOUSE-PETERS; CHANG, 2011). Pulido-Calvo et al. (2007) used feed-forward ANNs trained with Levenberg-Marquardt (LM) to model irrigation water demand. Adamowski and Karapataki (2010) compared the performance of three Multilayer Perceptron ANNs for predicting short-term water demand: LM, resilient backpropagation e conjugate gradient Powell-Beale.

Bennett, Stewart and Beal (2013) applied the neural network approach to develop a residential water end-use demand forecasting model. Firat, Yurdusev and Turan (2008) compared three ANN techniques: (Generalized Regression Neural Networks (GRNN), Feed Forward Neural Networks (FFNN) and Radial Basis Neural Networks (RBNN) to Multiple Linear Regression (MLR). ANN also outperformed regression and time-series models (BOUGADIS; ADAMOWSKI; DIDUCH, 2005; GHIASSI; ZIMBRA; SAIDANE, 2008).

ANNs are statistical models build through an iterative self-learning process. An ANN is a network of interconnected nodes structured as layers (input, hidden and output) with weighted connections (GHIASSI; ZIMBRA; SAIDANE, 2008). The weights vary according to the algorithm used. The network accumulates knowledge in each layer until the process behavior is captured (HOUSE-PETERS; CHANG, 2011).

The urban demand was projected with a Multilayer Perceptron (MLP) Neural Network and trained with Backpropagation of the error (RUMELHART; HINTON; WILLIAMS, 1986). An MLP network has at least three layers: input, output and hidden (Figure 3).

Figure 3 – MLP configuration.



Source: Adapted from Firat, Yurdusev and Turan (2008).

The sum of weighted input signals calculated by Equation 15 is transferred by a nonlinear activation function given in Equation 16. The response of network is compared with the observed values and the network error is calculated with Equation 17 (FIRAT; YURDUSEV; TURAN, 2008).

$$Y_{net} = \sum_{i=1}^{N} x_i w_i + w_0 \tag{15}$$

$$Y_{out} = f(Y_{net}) = \frac{1}{1+e^{-Y_{net}}} \tag{16}$$

$$J_r = \frac{1}{2}\sum_{i=1}^{k}(Y_{obs} - Y_{out})^2 \tag{17}$$

Where f(Y_{net}) is the nonlinear activation function, x_i is the neuron input, w_i is weight coefficient of each neuron input, w_0 is bias, Y_{out} is the system response, J_r is the error between observed and the network result and Y_{obs} is the observation value.

Determining the number of hidden layers is a difficult task an there is no consensus about it (REED; MARKSII, 1999). Usually, one or two hidden layers are usually enough to solve any nonlinear problem (LIPPMANN, 1987). For this reason, an MLP with two hidden layers was used in this research.

The number of nodes in the hidden layer is also hard to define and usually is function of the input and output layers size (BERRY; LINOFF, 1997, BOGER; GUTERMAN, 1997). However, other aspects should be considered, such as the neural network architecture and the training samples database (KARSOLIYA, 2012). After comparing ANN performance with different number of hidden layer neurons, the size was set as the number of explanatory variables (n) added by one. The size was the same for both hidden layers. Table 2 summarizes the architecture of the neural network.

Table 2 – Architecture of the ANN-MLP.

Parameter	Value
Number of layers	4
Number of hidden layers	2
Number of nodes in the hidden layers	n + 1
Training algorithm	Backpropagation
Hidden layer transfer function	Act Logistic
Iniatilization function	Randomize Weights
Number of iterations	1000

Source: Elaborated by the author.

The performance of ANN models for each variable subset was compared with the observations and the best model was identified. The complete data set was randomly divided into two sets: training (80%) and test (20%). The best fit models were also trained and tested by RF and PLSR.

4 RESULTS AND DISCUSSION

This section presents the variable subsets chosen by each method. One of them was used to compare three predictive models: ANN, PLSR and RF. After defining the best model, each subset was tested as input for the predictive models with best performance.

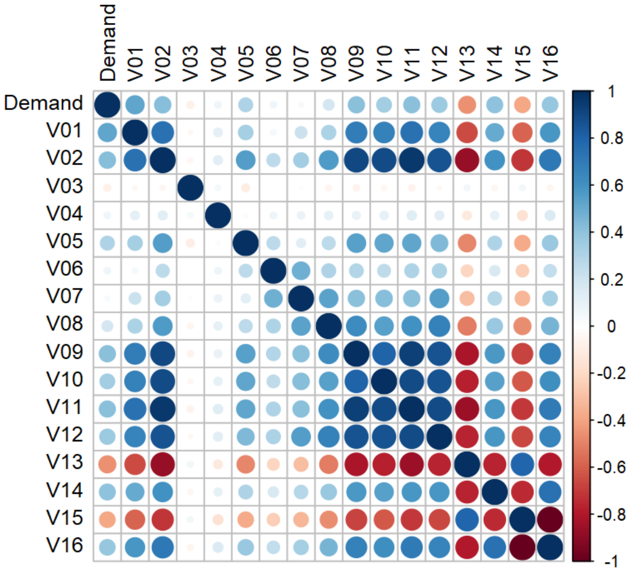
4.1 Variable selection

Seven methods of variable selection were compared. First, empirical decision based on Pearson correlation matrix is presented. Second, PLSR based methods (filter and wrapper) subsets are compared. Finally, the variable importance ranking obtained through RF serves as input for five different feature subsets.

4.1.1 Pearson correlation

The correlation matrix is illustrated in Figure 4. Except for employment rate (V03), Gini index (V04), percentage of 5 to 6 years old enrolled in school (V06), 6 to 14 years old enrolled in school (V07), percentage of 11 to 13 years old enrolled in final years of elementary school (V08), (r = -0.08, 0.07, 0.07, 0.04 and 0.18, respectively), all other variables are strongly associated with water consumption at the 0.05 significance level. Percentage of 1 to 14 years old (V13) and male residents (V15) are negatively correlated to water consumption.

Figure 4 – Correlation matrix.



Source: Elaborated by the author.

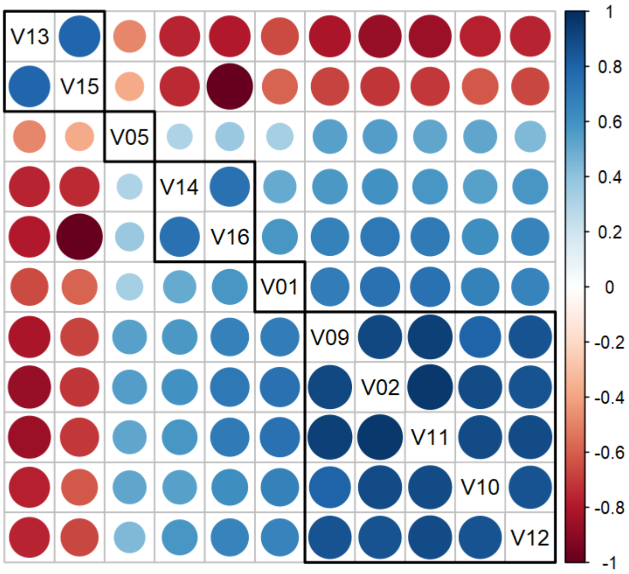
Independent variables are also correlated to each other, such as per capita income, associated with life expectancy at birth (r = 0.74), percentage of 18 to 20 years old that finished High School (r = 0.69) and expected years of schooling (r = 0.66). Besides, all education variables are correlated, suggesting that there could be a multicollinearity problem in the predictive model.

The variables with low correlation to water demand were removed before identifying the correlation clusters, presented in Figure 5. Per capita income (V01) and percentage of people in households with bathrooms and running water (V05) are isolated, indicating these variables might be important for water demand prediction.

Arbues e Villanua (2006) indicated that income has a direct relationship with the water demand. They also found that changes in income imply in a slow adaptation of water consumption, since residential users have well-established habits.

Hence, per capita income (V01) and percentage of people in households with bathrooms and running water (V05) were included in the empirical subset of variables.

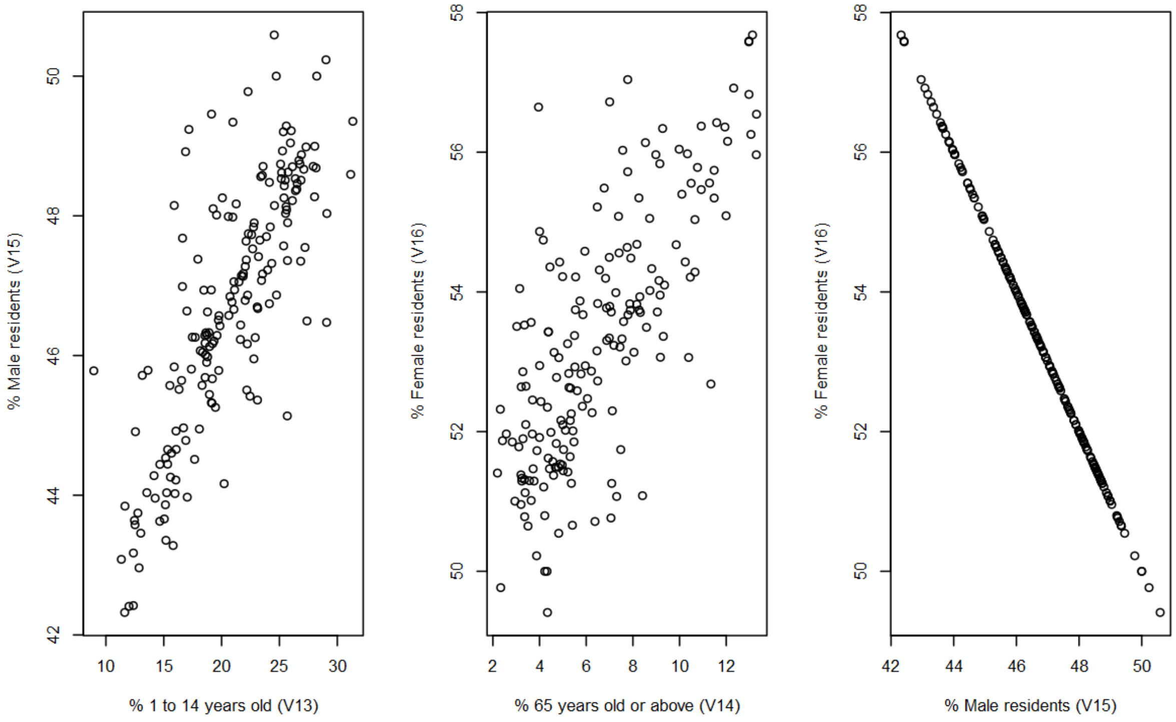
Figure 5 – Correlation clusters.



Source: Elaborated by the author.

Percentage of 1 to 14 years old (V13) and male residents (V15) as well as percentage of 65 years old or above (V14) and female residents (V16) are correlated and were aggregated in clusters. Figure 6 illustrates the correlation between V13 and V15 (r = 0.797), plot in the left; and V14 and V16 (r = 0.741), plot in the center. This could indicate they should not be in the same subset; however, they can have complementary information. On the other hand, variables perfectly correlated should not be included. This is the case of V15 and V16 (plot in the right), that have a negative correlation of r = -1. These variables are redundant and adding both will not imply in information gain for the model.

Figure 6 – Correlation between V13 and V15; V14 and V16; V15 and V16.



Source: Elaborated by the author.

Besides percentage of 1 to 14 years old (V13), female residents (V16) and 65 years or above (V14) were added in the selection, since they might have complementary information.

All variables in the bottom right cluster are correlated with each other. For example, V09 (% 18 to 20 years old that finished High School) is associated with V11 (% 18 or older that finished Elementary School; r = 0.935). This might be true because people who finished High School certainly completed Elementary School.

All education indices (V09, V10, V11 and V12) are strongly correlated to life expectancy at birth (V02), with Pearson correlation above 0.86. The existence of a positive association between life expectancy and education is well documented (HANSEN; STRULIK, 2017). The reverse causality (causal impact of education on health and longevity) is also addressed in micro-econometric literature (CUTLER; LLERAS-MUNEY; VOGL, 2008).

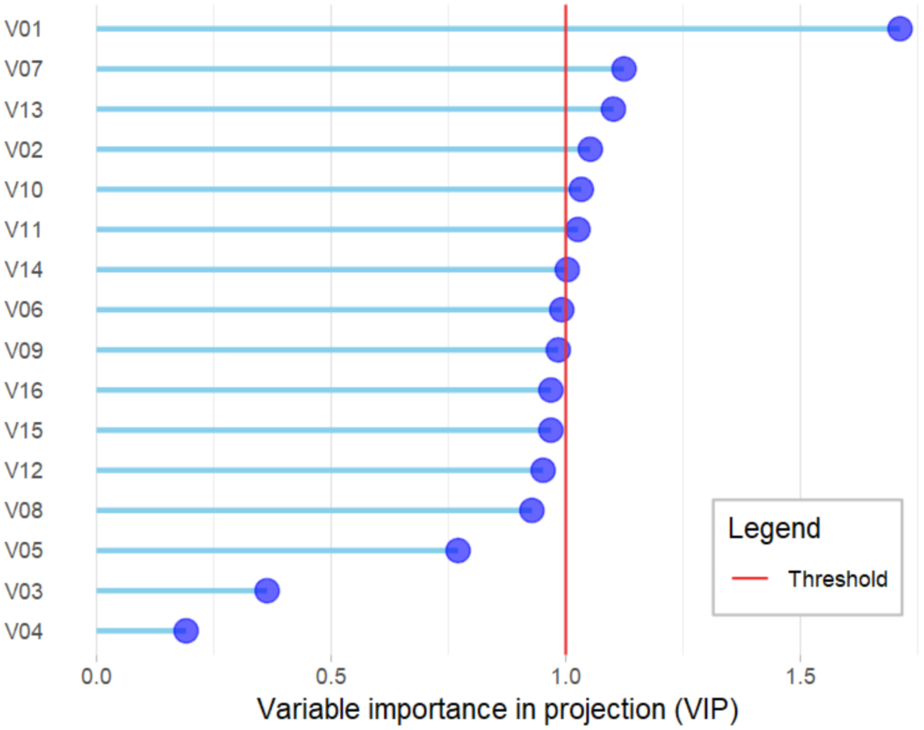
Education level was represented by the number of college-educated people in some studies (HOUSE-PETERS; PRATT; CHANG, 2010; SHANDAS; PARANDVASH, 2010). Shandas e Parandvash (2010) found that an increase of 100 college-educated residents per block group results in a 0.2 acre-foot reduction in water consumption. However, House-peters, Pratt e Chang (2010) found a positive association between education and water demand.

Among the information related to education, V09 (% 18 to 20 years old that finished High School) and V12 (expected years of schooling) were chosen to be part of the subset. Life expectancy (V02) was also added. The final empirical subset is presented at the end of this section.

4.1.2 Variable importance in projection (VIP)

The higher a variable's VIP score, the more influential it is in determining the predictive model outputs. Variables with VIP scores lower than 1 were assumed as unimportant. As shown in Figure 7, 7 of the 16 variables scored more than 1.

Figure 7 – Variable importance according to PLSR-VIP.



Source: Elaborated by the author.

The least and most important variables were Gini Index (V04) and per capita income (V01), respectively. Among the variables related to education that scored more than 1 are V07 % 6 to 14 years old enrolled in school (V07), % 15 to 17 years old that finished Elementary School (V10) and % 18 or older that finished Elementary School (V11). Basically, the ones related to primary education

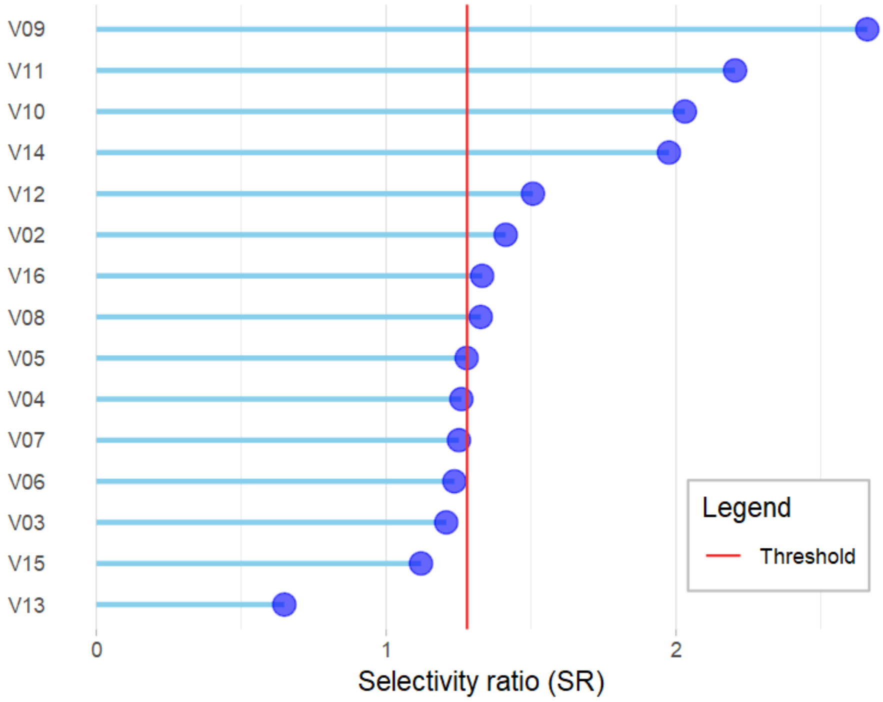
The method selected a variable with low correlation to water demand (V07), indicating irrelevant variation might have influenced VIP. However, a variable that seems to be useless by itself can provide a significant performance improvement when taken with others (GUYON; ELISSEEFF, 2003).

All other components of the subset are reasonable and bring different socioeconomic aspects of household. Per capita income (V01) representing economy; life expectancy (V02), health; V07, V10 and V11, education; % 1 to 14 years old (V13) and % 65 years or above (V14), demographic.

4.1.3 Selectivity ratio (SR)

In Figure 8, the features are ranked according to the SR values. The red line represents the threshold of 1.279, calculated through F-test F(0.05, 180, 179). Variable V01 (per capita income) was left out of the plot because its score is 10^3 times greater than others. The substantial SR value of V01 indicates a strong correlation between its predictive part and the demand and a low unexplained variance. 7 of the 16 variables scored less than 1.279.

Figure 8 – Variable importance according to SR.



Source: Elaborated by the author.

While male residents (V15) was left out of the subset, female residents (V16) was included. Percentage of elderly residents is also part of the subset. All variables related to education were selected, except for percentage of 5 to 6 years old (V06) and 6 to 14 years old enrolled in school (V07). Like VIP, SR selected a variable with low correlation to water demand (V08 – percentage of 11 to 13 years old enrolled in final years or that finished Elementary School).

4.1.4 Wrapper methods

Unlike VIP and SR, the wrapper methods present subsets rather than rankings of variables. Thus, they select features that work well together.

Employment rate (V03) and Gini index (V04) were not selected in any of them, while percentage of 11 to 13 years old enrolled in Elementary School (V08) was only present in the SwPA subset. The only method that included per capita income into the set was REP-PLS, despite its strong correlation to demand (r = 0.524).

BVE-PLS, method for elimination of uninformative variables, selected only 6 of the 16 features and did not include those with low correlation to water demand. UVE-PLS included four more variables to the BVE-PLS subset: percentage of people in households with bathrooms and running water (V05), percentage of 15 to 17 years old that finished Elementary School (V10), expected years of schooling (V12) and Male residents (V15).

SwPA-PLS and UVE-PLS selected larger subsets, both with 10 components. They were also the only methods to include all variables related to demographic aspects of households. BE-PLS and REP-PLS selected more parsimonious subsets, with 6 and 5 components respectively.

SwPA-PLS was the only method to include V08, that presents low correlation to demand (r = 0.18). Table 3 summarizes the variables selected by filter and wrapper methods based on Partial Least Squares Regression.

Table 3 – Variables selected by PLSR-based methods.

ID	Variable	VIP	SR	BVE-PLS	UVE-PLS	SwPA-PLS	REP-PLS
V01	Per capita income	*	*				*
V02	Life expectancy at birth	*	*	*	*	*	*
V03	Employment rate - 18 years old and older						
V04	Gini Index						
V05	% of people in households with bathrooms and running water				*	*	
V06	% 5 to 6 years old enrolled in school						
V07	% 6 to 14 years old enrolled in school	*					

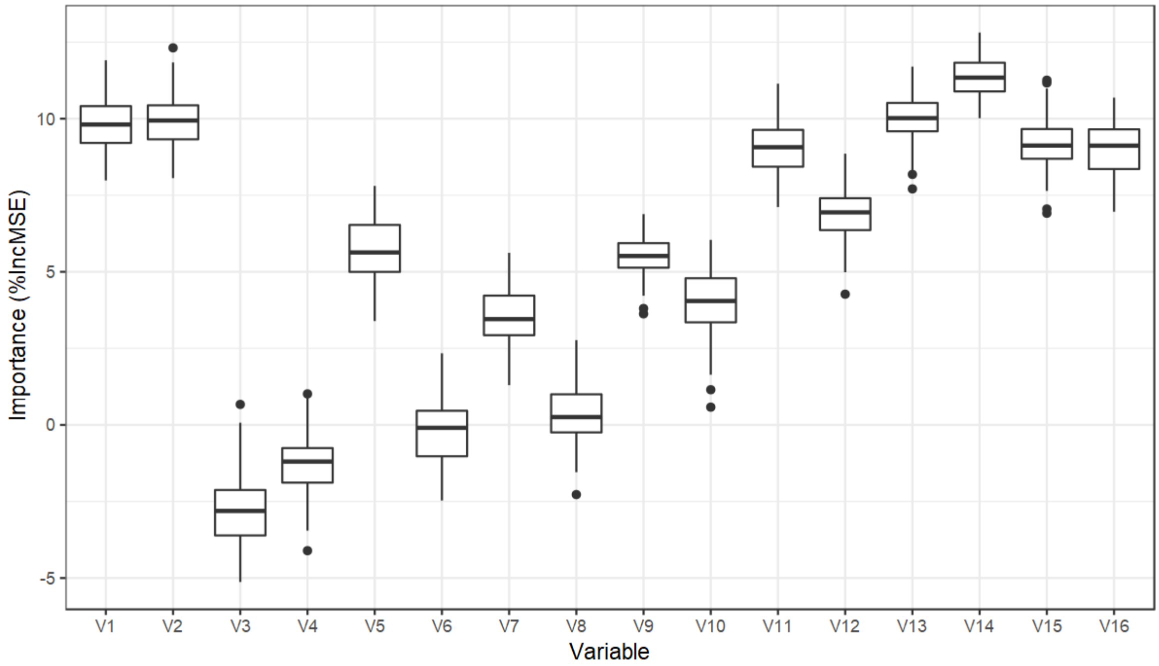
ID	Variable	VIP	SR	BVE-PLS	UVE-PLS	SwPA-PLS	REP-PLS
V08	% 11 to 13 years old enrolled in final years or that finished Elementary School		*			*	
V09	% 18 to 20 years old that finished High School		*	*	*	*	*
V10	% 15 to 17 years old that finished Elementary School	*	*		*	*	
V11	% 18 or older that finished Elementary School	*	*	*	*	*	
V12	Expected years of schooling		*		*		
V13	% 1 to 14 years old	*		*	*	*	*
V14	% 65 years old or above	*	*	*	*	*	*
V15	% Male residents				*	*	
V16	% Female residents		*	*	*	*	

Source: Elaborated by the author.

4.1.5 Random Forest

Figure 9 represents the box plot of the %IncMSE variation for 100 runs of the model.

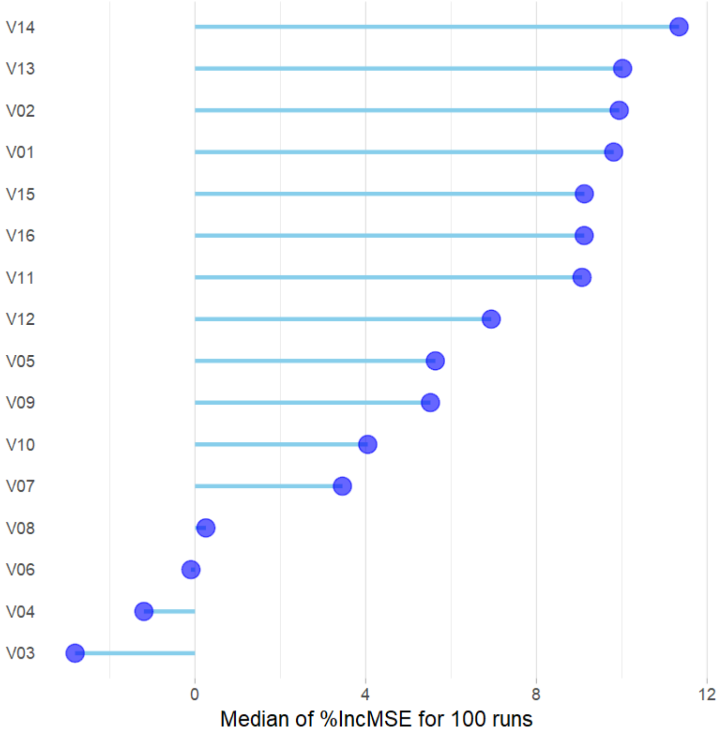
Figure 9 – Box plot of variable importance according to RF.



Source: Elaborated by the author.

The median of the importance measure for each variable was taken for ranking them. Variable importance ranking is presented in Figure 10.

Figure 10 – Variable importance according to RF.



Source: Elaborated by the author.

Percentage of 65 years old or above (V14) was classified as the most important variable, followed by percentage of 1 to 14 years old (V13). Employment rate (V03), Gini index (V04), percentage of 5 to 6 years old enrolled in school (V06) scored negative values of IncMSE. This indicates the random permuted variables worked better than the original (lower MSE), so they are not important for prediction. Besides, 11 to 13 years old enrolled in final years (V08) scored a low value of IncMSE, indicating it is also a bad predictor of water demand.

The cut-off to select features is arbitrary. In total, 5 models were developed based on the RF selection. The first subset is composed by the top 6 features and the other groups were formed by adding the subsequent variable with higher IncMSE. After, each subset was used as input for an ANN model and the correspondent RMSE, Nash and correlation were computed.

Table 4 presents the subsets based on RF (with 6 and 10 components) and empirical evidence. RF-6 is more parsimonious and does not contemplate socioeconomic aspects such as education. RF-10, on the other hand, includes variables related to demographic, education and housing characteristics.

Table 4 – Variables selected by RF and Empirical methods.

ID	Variable	RF-6	RF-9	RF-10	Empirical
V01	Per capita income	*	*	*	*
V02	Life expectancy at birth	*	*	*	*
V03	Employment rate - 18 years old and older				
V04	Gini Index				
V05	% of people in households with bathrooms and running water		*	*	*
V06	% 5 to 6 years old enrolled in school				
V07	% 6 to 14 years old enrolled in school				
V08	% 11 to 13 years old enrolled in final years or that finished Elementary School				
V09	% 18 to 20 years old that finished High School		*	*	*
V10	% 15 to 17 years old that finished Elementary School				
V11	% 18 or older that finished Elementary School		*	*	
V12	Expected years of schooling			*	*
V13	% 1 to 14 years old	*	*	*	*
V14	% 65 years old or above	*	*	*	*
V15	% Male residents	*	*	*	
V16	% Female residents	*	*	*	*

Source: Elaborated by the author.

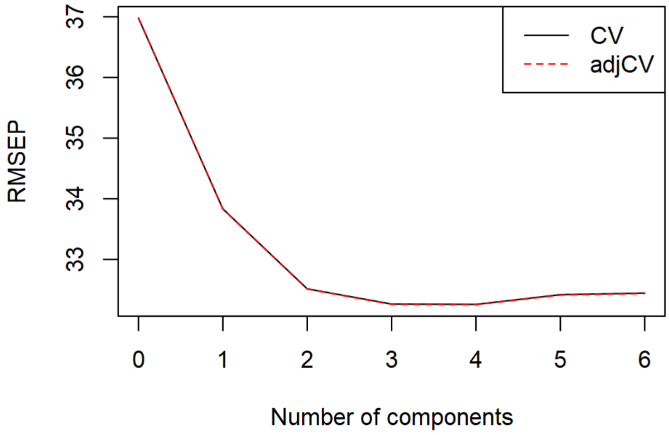
4.2 Calibration and validation of the predictive models

ANN and PLSR models were trained and tested with the same datasets. A statistical analysis of the outputs was performed, and the parameters were compared for both models. RF was also used for prediction and compared to ANN and PLSR.

4.2.1 PLSR

The variable subset selected by the VIP method was used to compare the prediction models. The first to be evaluated was the PLSR, initially set with six components. The result of the cross-validation (CV) of the fitted PLSR model is illustrated in Figure 11. The plot represents the Root Mean Squared Error of Prediction (RMSEP) as functions of the number of components, where “adjCV” is a bias-corrected CV estimate.

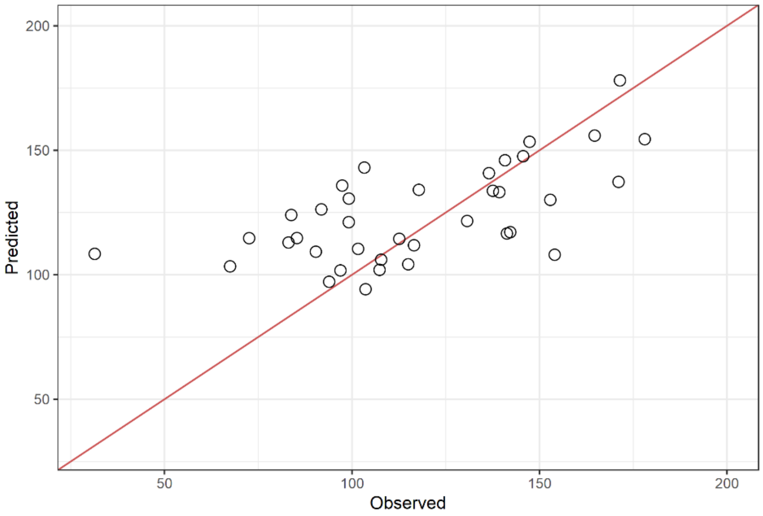
Figure 11 – Cross-validated RMSEP curves.



Source: Elaborated by the author.

Although four components give the smallest RMSE value, three components are enough, since the RMSE reduction is small. The prediction with three components versus observed values for the test set are plotted in Figure 12. The plot indicates the predictive model was mostly overestimating the actual output, considering most of observations are above the target line.

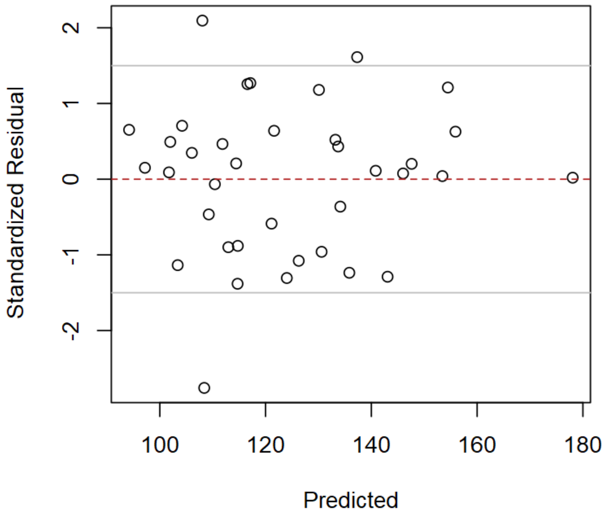
Figure 12 – Predictions for the PLSR model.



Source: Elaborated by the author.

Figure 13 presents the residual plot for the PLSR model, where residual is the difference between observed and predicted values. The residual was rescaled to have a mean of zero and a standard deviation of one. The values are symmetrically distributed, variating between -1.5 and 1.5 (grey lines). This indicates the data is probably homoscedastic, since different residuals appear to have similar variances.

Figure 13 – Residual plot for the PLSR model.

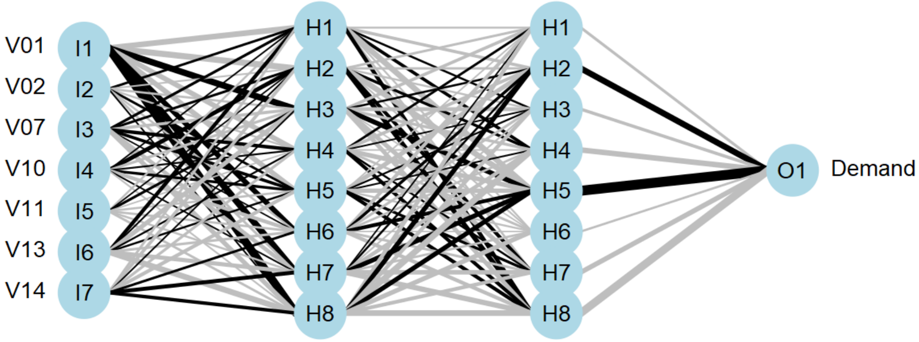


Source: Elaborated by the author.

4.2.2 Artificial Neural Network

Figure 14 represents a neural interpretation diagram as in Özesmi and Özesmi (1999). Black and grey lines represent positive weights between layers and negative weights, respectively. Line thickness proportional to relative magnitude of each weight. The first layer is the input layer (the number of nodes is equal to the input variables). The hidden layers are plotted with each node in each layer labelled as H. The output layer, with the node labeled as O1, is the water demand.

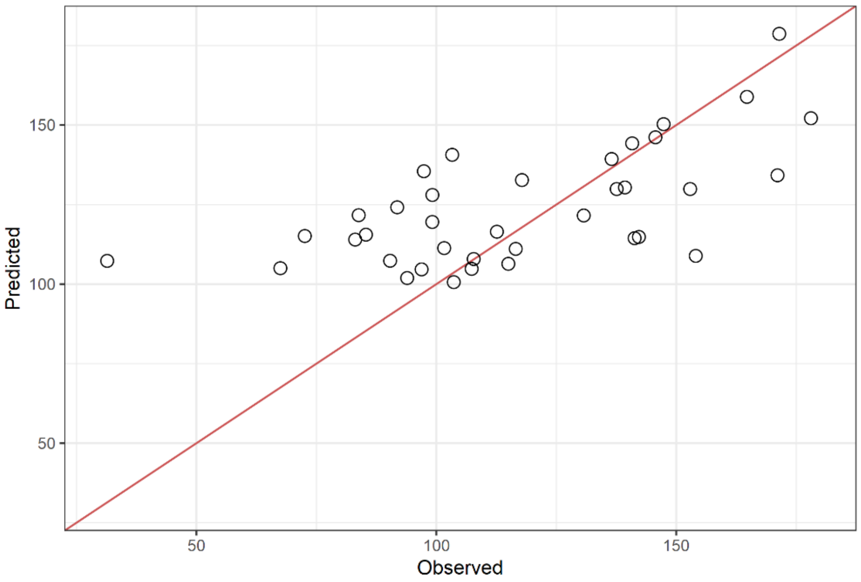
Figure 14 – Neural interpretation diagram.



Source: Elaborated by the author.

The prediction with the ANN versus observed values for the test set are plotted in Figure 15. The plot indicates the predictive model provided a reasonable estimation, since part of the observations are close to the target line. However, for low values of water demand, the model tends to overestimate the predictions. A few predictions are underestimated, especially for observed value over 150 Lpd^{-1}. This might be due to the presence of HDUs with a wide gap of water consumption in the training set.

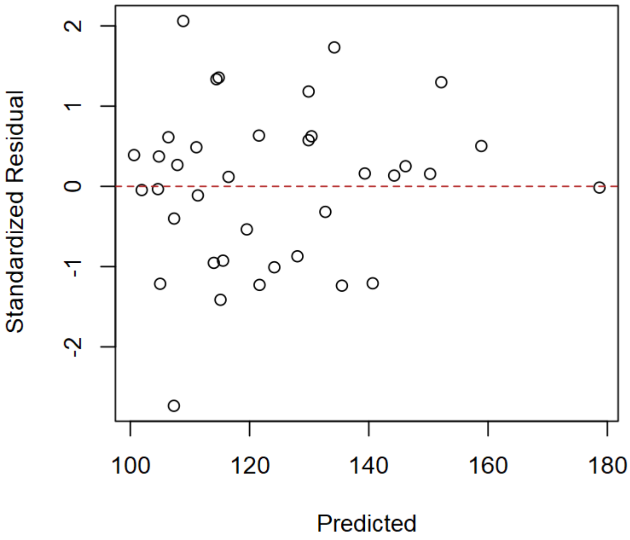
Figure 15 – Predictions for the ANN model.



Source: Elaborated by the author.

Figure 16 presents the residual plot for the ANN model. Like the PLSR model, the data is symmetrically distributed, indicating similar residual’s variance. There are also overestimated and underestimated values among the observations.

Figure 16 – Residual plot for the ANN model.

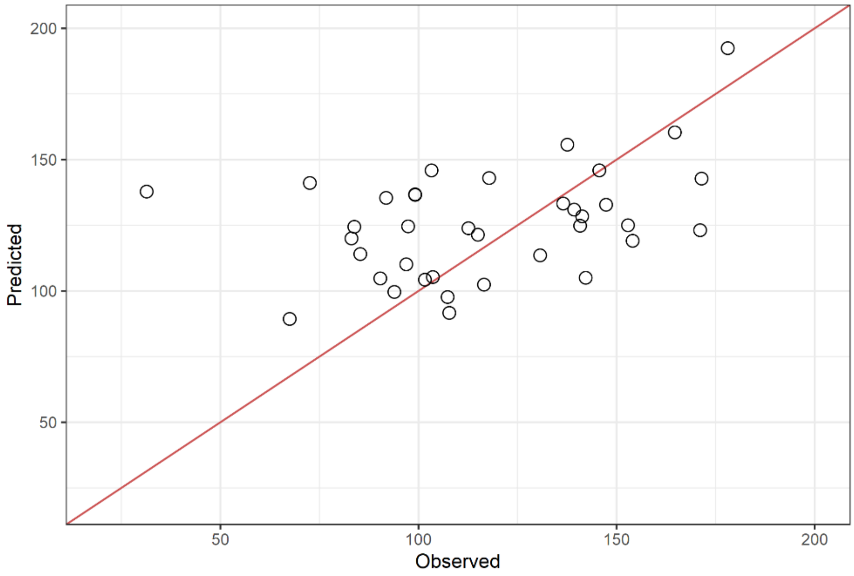


Source: Elaborated by the author.

4.2.3 Random Forest

The predicted values of the RF model versus observed values for the test set are plotted in Figure 17. The plot illustrates overestimated and underestimated values of water demand, with a few outliers. The observations are more dispersed around the target line than ANN and PLSR’s.

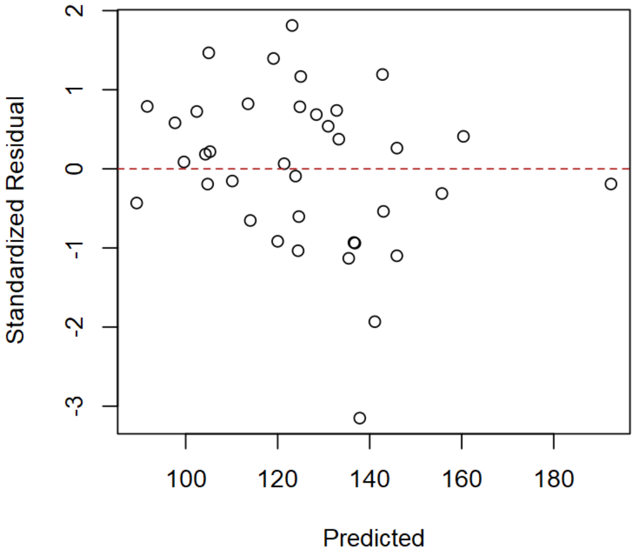
Figure 17 – Predictions for the RF model.



Source: Elaborated by the author.

Figure 18 presents the residual plot for the RF model. The plot indicates the residual’s variance is probably not uniform.

Figure 18 – Residual plot for the RF model.



Source: Elaborated by the author.

Table 5 presents parameter values for the predictive models. ANN and PLSR had very similar performances, while RF resulted in low NSE and correlation values, as well as greater values of RMSE in both training and test.

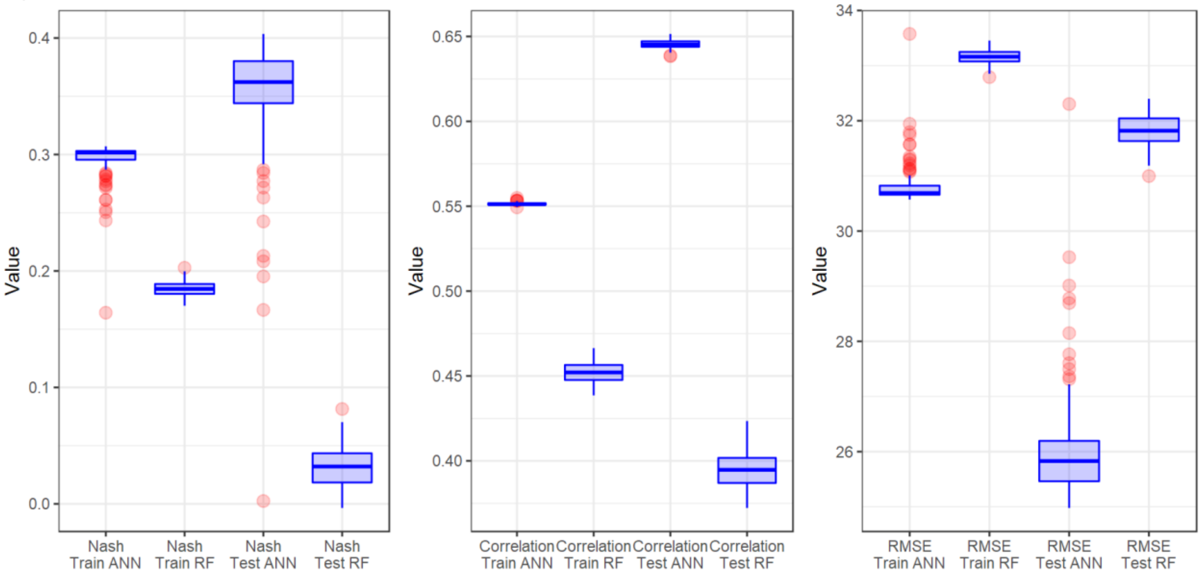
Table 5 – Prediction model performance.

	Train				Test			
Model	NSE	RMSE	Correlation	NRMSE	NSE	RMSE	Correlation	NRMSE
PLSR	0.303	30.652	0.551	0.253	0.354	25.994	0.636	0.222
ANN	0.304	30.637	0.552	0.253	0.362	25.836	0.644	0.221
RF	0.174	33.365	0.447	0.275	0.032	31.820	0.397	0.272

Source: Elaborated by the author.

Figure 19 represents performance comparison between RF and ANN models for predicting water demand. The models were tested using the variables selected by VIP method. All parameters indicate the ANN model had a better performance than RF. ANN presented greater NSE and correlation values and smaller RMSE than RF during train and testing. Therefore, while RF might be appropriated for variable selection, it is not suitable for prediction of future water demand.

Figure 19 – RF and ANN performance comparison.



Source: Elaborated by the author.

4.3 Performance of feature subsets

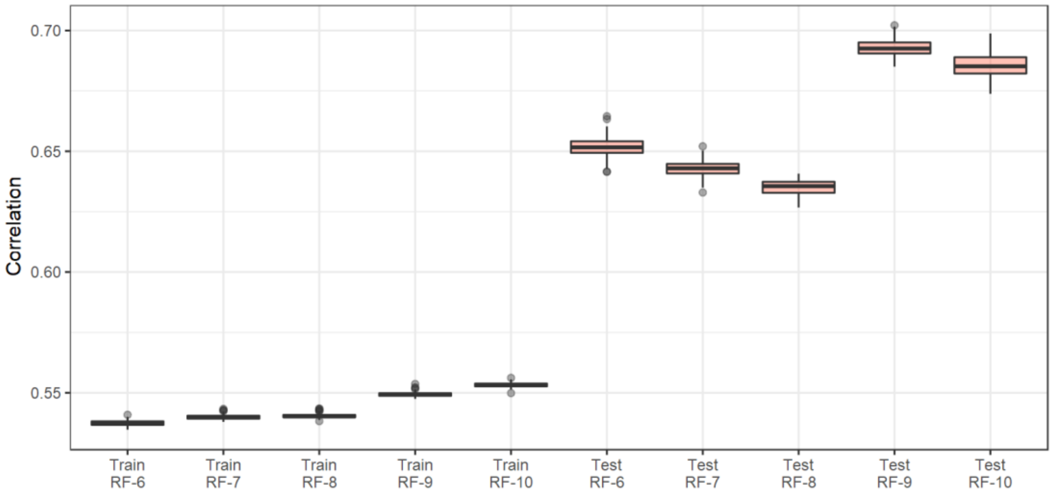
In this section, each variable subset was used as input for the ANN model. The RMSE, Pearson Correlation and Nash-Sutcliffe values of each model were compared to choose the best subset.

4.3.1 Performance of RF feature subsets (ANN model)

Before comparing the selection methods, the subsets chosen with RF were compared to define the best of them. Each group of 6 to 10 variables was used as input to the ANN prediction model. Figure 20, Figure 21 and Figure 22 present boxplots of correlation, Nash and RMSE for each subset.

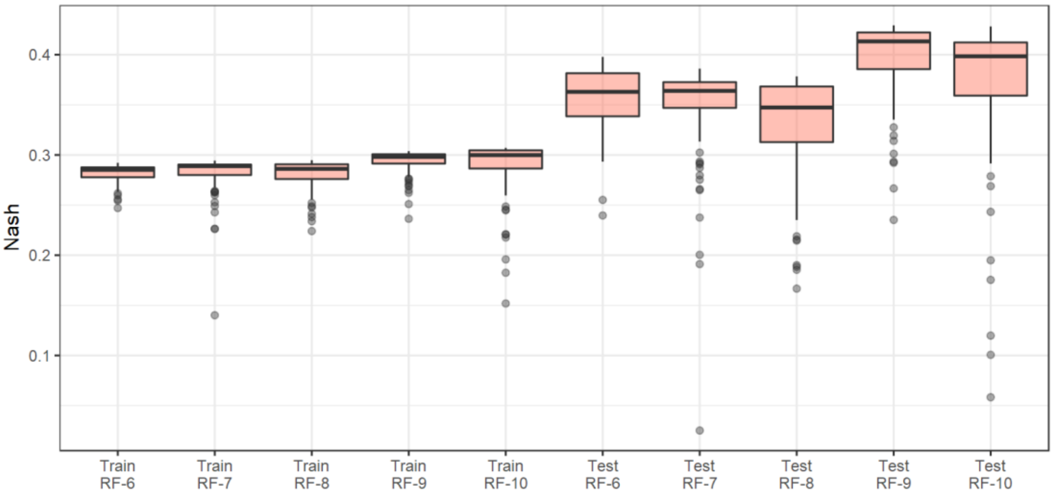
The subset with nine variables is the best of all inputs, achieving better correlation and Nash values than other subsets and a lower RMSE. RF-9 subset includes variables representing different socioeconomic aspects. The only difference between RF-9 and RF-10 is the variable Expected years of schooling (V12), indicating it might not add important information to the subset, since RF-10 had a similar performance to RF-9.

Figure 20 – Performance of RF subsets (Correlation).



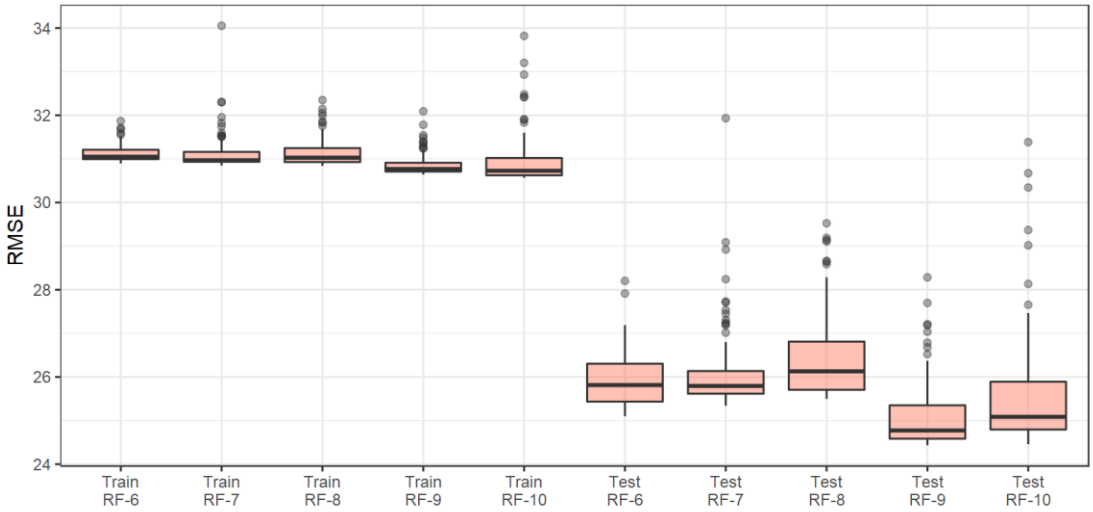
Source: Elaborated by the author.

Figure 21 – Performance of RF subsets (NSE).



Source: Elaborated by the author.

Figure 22 – Performance of RF subsets (RMSE).

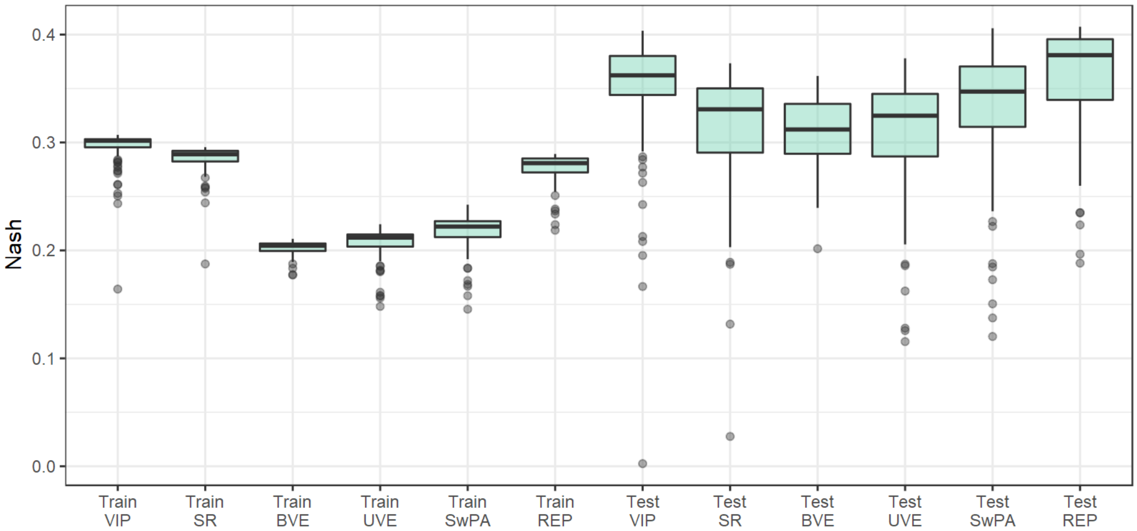


Source: Elaborated by the author.

4.3.2 Performance of PLSR-based feature subsets (ANN model)

Figure 23 represents a comparison between each variable subset selected with the PLSR-based methods presented before. When variables selected by VIP and REP methods were used as input for the ANN model, the best Nash values were achieved. The results imply these methods might be a good choice for variable selection.

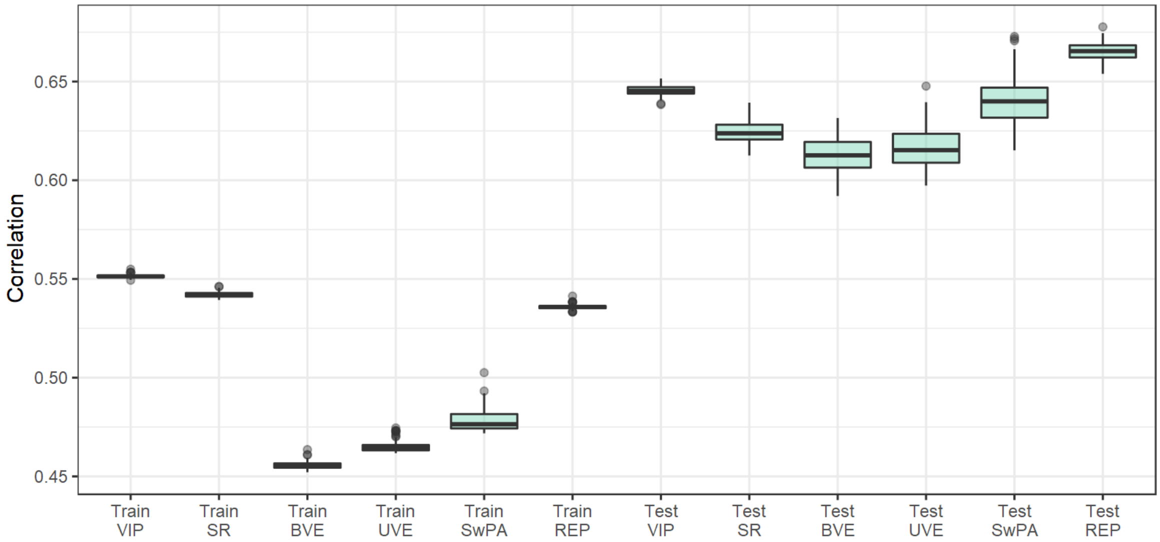
Figure 23 – Performance of PLSR-based variable selection methods (NSE).



Source: Elaborated by the author.

The variable subsets selected by BVE and UVE resulted in the least satisfactory performance. These are among the methods that exclude per capita income of their subsets. Besides, UVE included two perfectly correlated variables.

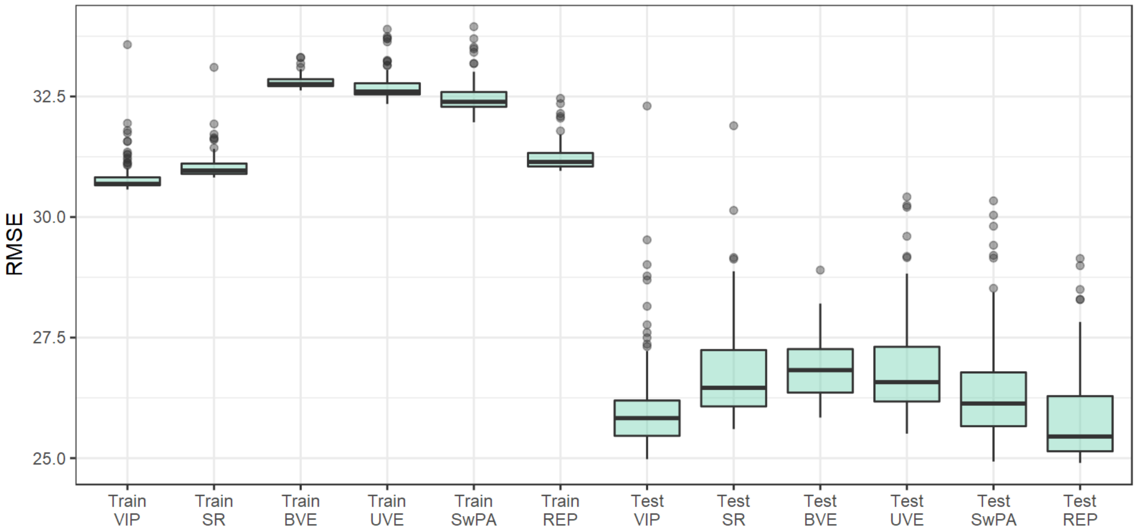
Figure 24 – Performance of PLSR-based variable selection methods (Correlation).



Source: Elaborated by the author.

In terms of RMSE, VIP and REP had a slightly better performance than other classification methods (Figure 25). They achieved the lowest values during training and test. BE, UVE and SwPA scored a high RMSE during train.

Figure 25 – Performance of PLSR-based variable selection methods (RMSE).

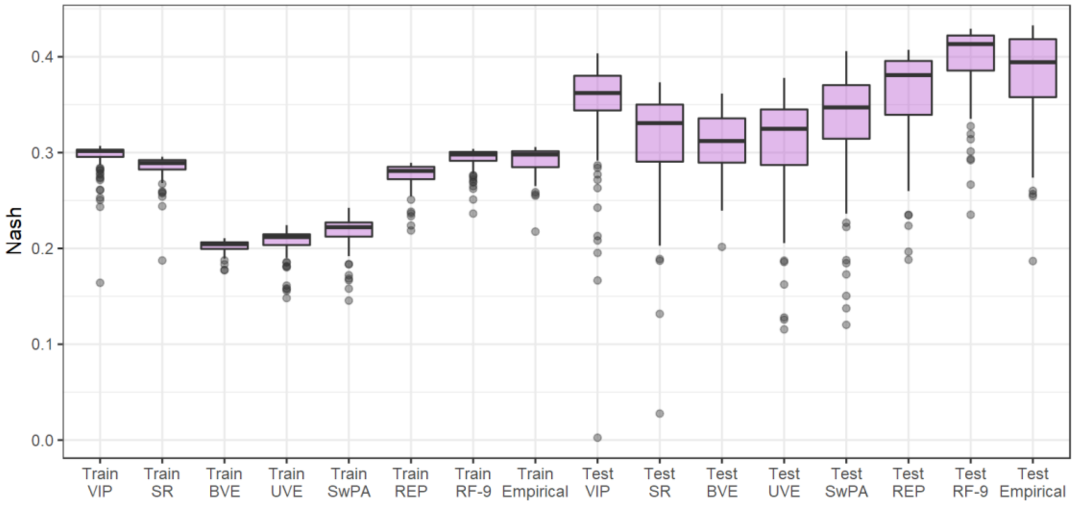


Source: Elaborated by the author.

4.3.3 Performance of all feature subsets (ANN model)

Figure 26 represents a comparison of NSE values between the variable subset for all the methods presented here. The prediction model performed very well when the RF-9 variable set was taken as input, with a Nash value above 0.4. The empirical subset also led to improvement in the predictive model.

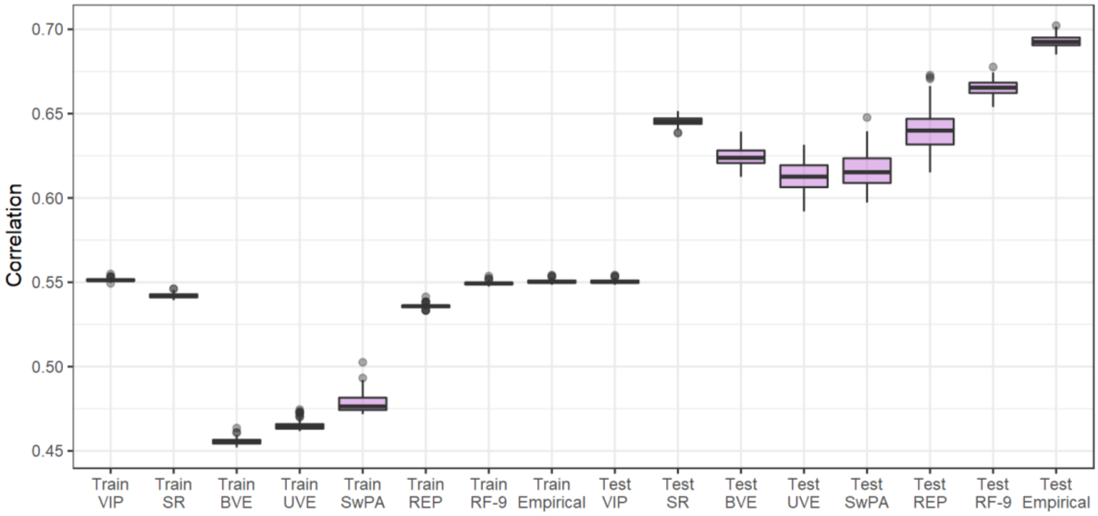
Figure 26 – Performance comparison of variable subsets for ANN (NSE).



Source: Elaborated by the author.

When evaluating correlation between observed and predicted values during test (Figure 27), SR, REP, RF-9 and Empirical subsets were the best choices for variable selection. During train and test demand values predicted with BVE, UVE and SwPA choices had the lowest correlation with observed values. Figure 27 represents the boxplots of correlation values for 100 runs of the ANN model for all the subsets.

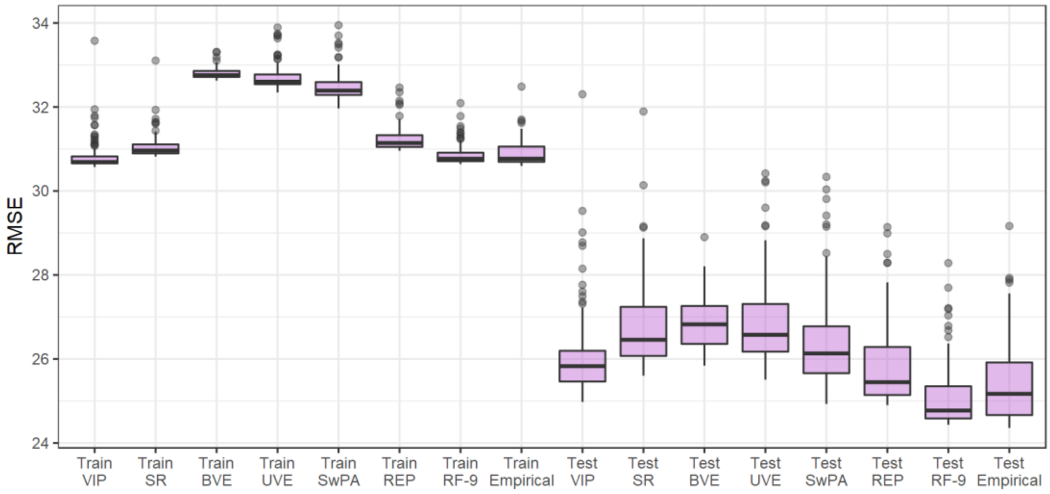
Figure 27 – Performance comparison of variable subsets for ANN (Correlation).



Source: Elaborated by the author.

Except for BVE, UVE and SwPA, all the subsets had similar performances regarding RMSE values of the prediction model during train. However, VIP, RF-9 and Empirical were slightly better. During test, RF-9 and Empirical resulted in the best subsets. SR is also a good method for variable selection. Figure 28 represents the boxplots of RMSE values for 100 runs of the ANN model for all the subsets.

Figure 28 – Performance comparison of variable subsets for ANN (RMSE).

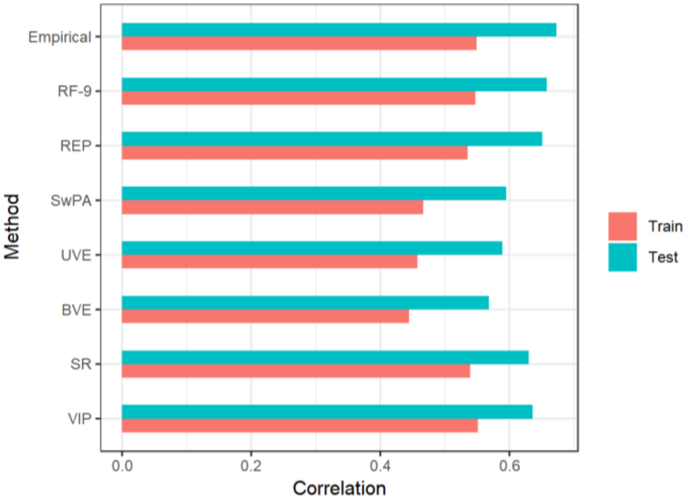


Source: Elaborated by the author.

4.3.4 Performance of all feature subsets (PLSR model)

The subsets were also compared using the PLSR model for prediction. Overall, they had very similar performances, but VIP, REP, RF-9 and Empirical resulted in better predictions. Figure 29 presents the correlation between observed and predicted values for each subset.

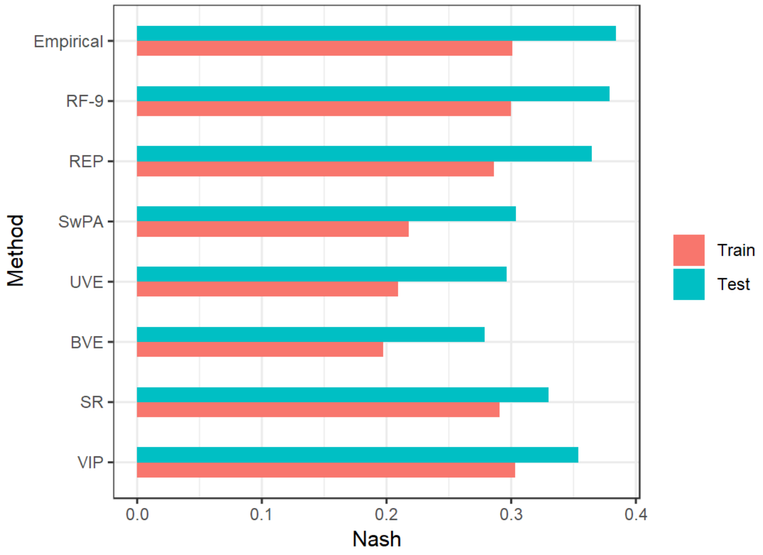
Figure 29 – Performance comparison of variable subsets for PLSR (Correlation).



Source: Elaborated by the author.

Figure 30 presents the NSE for the PLSR each subset. SwPA, UVE and BVE had the worst performances. VIP, REP, RF-9 and Empirical achieved greater NSE values than other subsets.

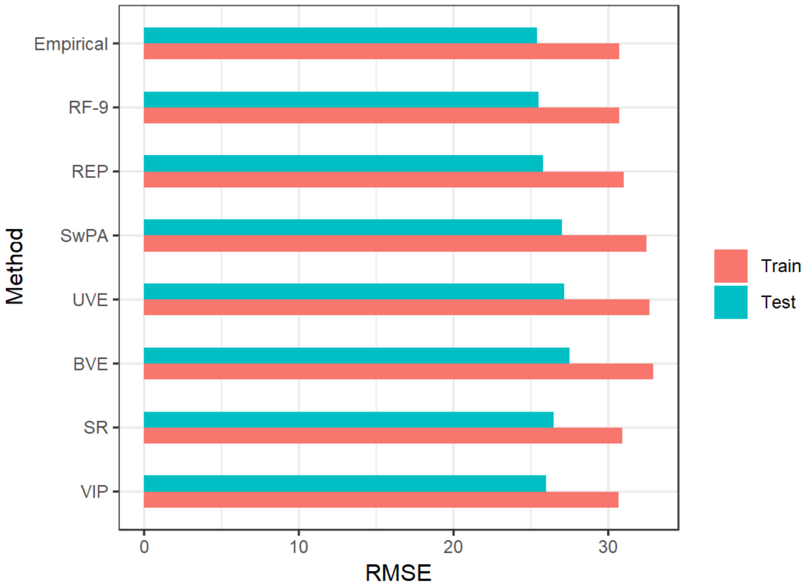
Figure 30 – Performance comparison of variable subsets for PLSR (NSE).



Source: Elaborated by the author.

Figure 31 represents the bar plot of the RMSE for each subset. Again, VIP, REP, RF-9 and Empirical had a slightly better performance among all subsets, with lower RMSE values.

Figure 31 – Performance comparison of variable subsets for PLSR (RMSE).



Source: Elaborated by the author.

4.3.5 Performance PLSR and ANN with best subsets

After evaluating the subsets selected by each method, the best choice was used to compare PLSR and ANN methods. Table 6 presents the parameters of each model when RF-9 and REP subsets were used as inputs for ANN and PLSR, respectively.

Table 6 – ANN and PLSR performance with their best subsets.

	Train				Test			
Model	NSE	RMSE	Correlation	NRMSE	NSE	RMSE	Correlation	NRMSE
PLSR	0.286	31.025	0.535	0.256	0.365	25.779	0.651	0.220
ANN	0.303	30.668	0.551	0.253	0.413	24.776	0.691	0.212

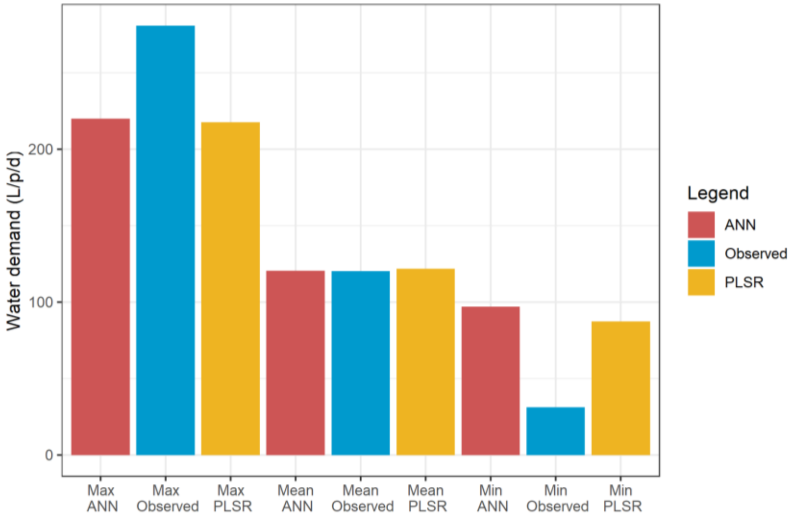
Source: Elaborated by the author.

Although VIP subset had a better performance than REP, this method was chosen because it is more parsimonious, containing only five variables while VIP has seven. For this reason, REP subset would be easier to estimate when constructing socioeconomic scenarios. Although larger, RF-9 subset had a better performance than other subsets. Besides, including a larger subset allowed an analysis if whether to include variables in the subset provides information and accuracy gain or not.

Both models had very similar performance for all parameters. Although a RMSE of 24 Lpd^{-1} might indicate the models are inaccurate, they still provide a better estimate than taking aa absolute value for all HDUs.

Figure 32 represents the maximum, minimum and mean of water demand obtained with ANN and PLSR and the real values. Mean of predicted values was very close to the observed demand. However, the maximum observed value was underestimated by both models, while the minimum was overestimated.

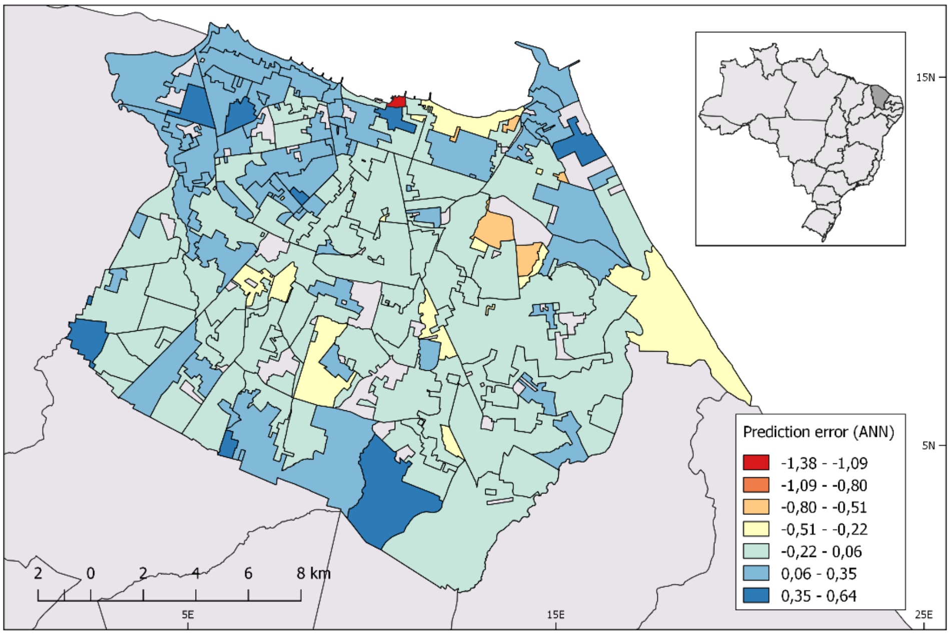
Figure 32 – Maximum, minimum and mean values of water demand.



Source: Elaborated by the author.

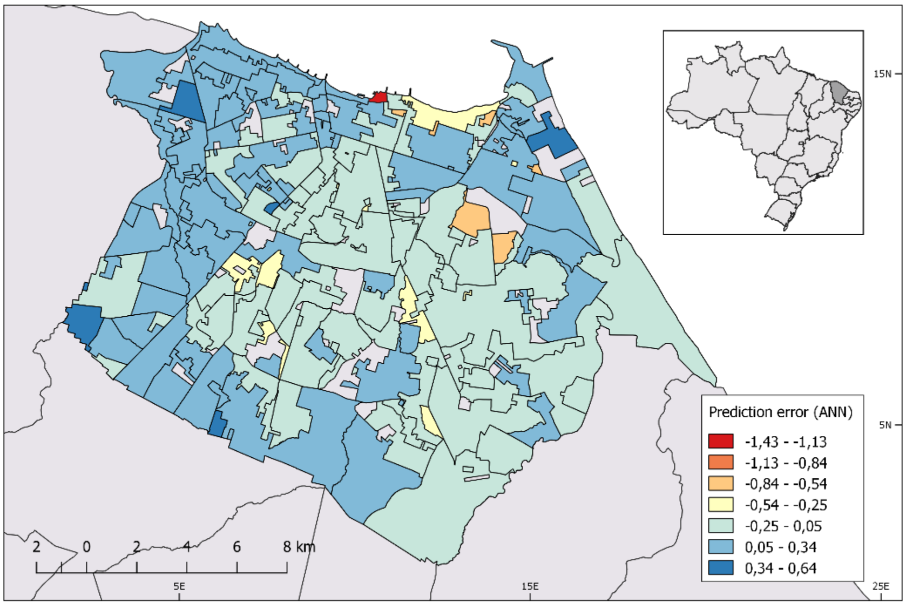
The prediction error for each HDU was estimated and are indicated in a map to allow a spatial analysis of the model's accuracy. Here, the prediction error corresponds to the difference between predicted and observed values divided by the mean of observed values. Figure 33 represents the error of the PLSR model estimations. For most HDUs, PLSR overestimates water demand (blue and green regions). The HDUs with the lowest consumption had overestimated predictions (dark blue), while HDUs with high water demand were underestimated (red and orange regions). The same pattern is observed for the ANN model, presented in Figure 34.

Figure 33 – Prediction error with PLSR for each HDU.



Source: Elaborated by the author.

Figure 34 – Prediction error with ANN for each HDU.



Source: Elaborated by the author.

4 CONCLUSIONS

Fortaleza presents heterogenous socioeconomic characteristics, implying in a spatially variable water demand. In a long-term perspective, changes in per capita income, education and demographic might influence consumption patterns. Thus, an accurate predictive model must include features that best represent these aspects. This study selected socioeconomic variables to forecast urban water demand with machine learning techniques and a linear model.

After collecting Census data, eight different approaches were used for variable selection, including filter, wrapper and embedded methods. Filter methods ranked per capita income and education as the most important variables, while RF considered education and life expectancy more important than profit.

VIP, REP and RF resulted in the best subsets for water demand prediction. Empirically chosen variables were also satisfactory predictors. RF and VIP selected a larger subset of variables, while REP resulted in a more parsimonious choice. All these methods included per capita income, life expectancy at birth, household composition variables (percentage of elderly and children) and an education related variable.

The percentage of 18 to 20 years old with a High School diploma was a good representation of education level. Although employment rate was included by some researchers into their predictive models (BRADLEY, 2004; KOO et al., 2005), any of the selection methods included this variable as an input for the model. In conclusion, the best subset choice provides enough information about consumers behavior and composition, while maintaining forecast quality.

Most predictive models include temperature, rainfall, water price and housing characteristics as explanatory variables (GOODCHILD, 2003; MOHAMED; AL-MUALLA, 2010; POLEBITSKI; PALMER, 2010). These were not part of this study, that focused on socioeconomic aspects of population. However, including weather related variables and water price might increase model performance.

In order to estimate future water demand, scenarios of economic growth must be built. One suggestion is to consider three possibilities: prevalence of the current conditions; growing economy with low efficiency in water resources management (pessimist) and growing economy with effective government management (optimistic). Tendencies of population aging and increase in life expectancy must also be considered.

Regarding the predictive model technique, ANN and PLSR outperformed RF. However, both models overestimated water demand for HDUs with lowest observed consumption and underestimated units of elevated water demand. ANN had better performance with variables selected by RF, while PLSR performed well with REP and VIP (both PLS-based methods).

Linear water demand functions are often chosen and have outperformed non-linear methods (ADAMOWSKI; KARAPATAK, 2010; DUERR et al.; 2018). However, considering a linear function implies that the change in water demanded in response to a social or economic change is the same at every socioeconomic level. Besides, other studies have proved ANN models are better than linear regression (JAIN; VARSHNEY; JOSHI, 2001; BOUGADIS, J.; ADAMOWSKI, K.; DIDUCH; 2005)

HDU level data allowed for increased understanding of how socioeconomic characteristics of household influence water consumption. Spatially aggregating estimates of water consumption allow the description of regional water use variability and further analysis of the consumers behavior.

Although data is spatially disaggregated, the model take the input as random observations across the study area. Therefore, this approach is not able to explain the influence of neighborhood characteristics on water consumption. However, the model still provides information for the development of conservation measures and policies as well as an expansion strategy.

REFERENCES

ADAMOWSKI, J.; KARAPATAKI, C. Comparison of multivariate regression and artificial neural networks for peak urban water-demand forecasting: Evaluation of different ANN learning algorithms. **Journal of Hydrologic Engineering**, v. 15, n. 10, p. 729-743, 2010.

ALHUMOUD, J. M. Freshwater consumption in Kuwait: analysis and forecasting. **Journal of Water Supply**: Research and Technology-Aqua, [s.l.], v. 57, n. 4, p. 279-288, jun. 2008. IWA Publishing. http://dx.doi.org/10.2166/aqua.2008.036.

ARBUES, F.; VILLANUA, I. Potential for Pricing Policies in Water Resource Management: Estimation of Urban Residential Water Demand in Zaragoza, Spain. **Urban Studies**, [s.l.], v. 43, n. 13, p. 2421-2442, dez. 2006. SAGE Publications. http://dx.doi.org/10.1080/00420980601038255.

AKARACHANTACHOTE, N.; CHADCHAM, S.; SAITHANU, K. Cutoff threshold of variable importance in projection for variable selection. **International Journal of Pure and Apllied Mathematics**, [s.l.], v. 94, n. 3, p. 1-16, 17 jul. 2014. Academic Publications. http://dx.doi.org/10.12732/ijpam.v94i3.2.

BENNETT, C.; STEWART, R. A.; BEAL, C. D. ANN-based residential water end-use demand forecasting model. **Expert Systems with Applications**, [s.l.], v. 40, n. 4, p. 1014-1023, mar. 2013. Elsevier BV. http://dx.doi.org/10.1016/j.eswa.2012.08.012.

BERRY, M. J.; LINOFF, G. Data Mining Techniques: For Marketing, Sales, and Customer Support. 1997.

BEHBOUDIAN, S. et al. A long-term prediction of domestic water demand using preprocessing in artificial neural network. **Journal of Water Supply**: Research and Technology-Aqua, [s.l.], v. 63, n. 1, p. 31-42, 17 out. 2013. IWA Publishing. http://dx.doi.org/10.2166/aqua.2013.085.

BOGER, Z.; GUTERMAN, H. Knowledge extraction from artificial neural network models. **IEEE International Conference on Systems, Man, And Cybernetics. Computational Cybernetics and Simulation**, [s.l.], p. 3030-3035, 1997. IEEE. http://dx.doi.org/10.1109/icsmc.1997.633051.

BOUGADIS, J.; ADAMOWSKI, K.; DIDUCH, R. Short-term municipal water demand forecasting. **Hydrological Processes**, [s.l.], v. 19, n. 1, p. 137-148, jan. 2005. Wiley. http://dx.doi.org/10.1002/hyp.5763.

BRADLEY, R. M. Forecasting Domestic Water Use in Rapidly Urbanizing Areas in Asia. **Journal of Environmental Engineering**, [s.l.], v. 130, n. 4, p. 465-471, abr. 2004. American Society of Civil Engineers (ASCE). http://dx.doi.org/10.1061/(asce)0733-9372(2004)130:4(465).

BREIMAN, L. Random Forests. **Machine Learning**, [s.l.], v. 45, n. 1, p. 5-32, 2001. Springer Nature. http://dx.doi.org/10.1023/a:1010933404324.

BREKKE, L. et al. Suburban Water Demand Modeling Using Stepwise Regression. **Journal - American Water Works Association**, [s.l.], v. 94, n. 10, p. 65-75, out. 2002. Wiley. http://dx.doi.org/10.1002/j.1551-8833.2002.tb09558.x.

BLUM, A. L.; LANGLEY, P. Selection of relevant features and examples in machine learning. **Artificial Intelligence**, [s.l.], v. 97, n. 1-2, p. 245-271, dez. 1997. Elsevier BV. http://dx.doi.org/10.1016/s0004-3702(97)00063-5.

CENTNER, V. et al. Elimination of Uninformative Variables for Multivariate Calibration. **Analytical Chemistry**, [s.l.], v. 68, n. 21, p. 3851-3858, jan. 1996. American Chemical Society (ACS). http://dx.doi.org/10.1021/ac960321m.

CUTLER, D. R. et al. Random forests for classification in ecology. **Ecology**, [s.l.], v. 88, n. 11, p. 2783-2792, nov. 2007. Wiley. http://dx.doi.org/10.1890/07-0539.1.

CUTLER, D. M.; LLERAS-MUNEY, A.; VOGL, T. **Socioeconomic status and health: dimensions and mechanisms**. National Bureau of Economic Research, 2008.

CHANG, H.; PARANDVASH, G. H.; SHANDAS, V. Spatial Variations of Single-Family Residential Water Consumption in Portland, Oregon. **Urban Geography**, [s.l.], v. 31, n. 7, p. 953-972, out. 2010. Informa UK Limited. http://dx.doi.org/10.2747/0272-3638.31.7.953.

CHONG, I.; JUN, C. Performance of some variable selection methods when multicollinearity is present. **Chemometrics And Intelligent Laboratory Systems**, [s.l.], v. 78, n. 1-2, p. 103-112, jul. 2005. Elsevier BV. http://dx.doi.org/10.1016/j.chemolab.2004.12.011.

DUERR, I. et al. Forecasting urban household water demand with statistical and machine learning methods using large space-time data: A Comparative study. **Environmental Modelling & Software**, [s.l.], v. 102, p. 29-38, abr. 2018. Elsevier BV. http://dx.doi.org/10.1016/j.envsoft.2018.01.002.

FARRÉS, M. et al. Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. **Journal of Chemometrics**, [s.l.], v. 29, n. 10, p. 528-536, 21 jul. 2015. Wiley. http://dx.doi.org/10.1002/cem.2736.

FIRAT, M.; YURDUSEV, M. A.; TURAN, M. E. Evaluation of Artificial Neural Network Techniques for Municipal Water Consumption Modeling. **Water Resources Management**, [s.l.], v. 23, n. 4, p. 617-632, 28 jun. 2008. Springer Nature. http://dx.doi.org/10.1007/s11269-008-9291-3.

FRANK, I. E. Intermediate least squares regression method. **Chemometrics And Intelligent Laboratory Systems**, [s.l.], v. 1, n. 3, p. 233-242, jul. 1987. Elsevier BV. http://dx.doi.org/10.1016/0169-7439(87)80067-9.

GARDINER, V.; HERRINGTON, P. **Water demand forecasting**. CRC Press, 2014.

GATO, S.; JAYASURIYA, N.; ROBERTS, P. Forecasting Residential Water Demand: Case Study. **Journal of Water Resources Planning and Management**, [s.l.], v. 133, n. 4, p. 309-

319, jul. 2007. American Society of Civil Engineers (ASCE). http://dx.doi.org/10.1061/(asce)0733-9496(2007)133:4(309).

GAUDIN, S. Effect of price information on residential water demand. **Applied Economics**, [s.l.], v. 38, n. 4, p. 383-393, 10 mar. 2006. Informa UK Limited. http://dx.doi.org/10.1080/00036840500397499.

GENUER, R.; POGGI, J.; TULEAU-MALOT, C. Variable selection using random forests. **Pattern Recognition Letters**, [s.l.], v. 31, n. 14, p. 2225-2236, out. 2010. Elsevier BV. http://dx.doi.org/10.1016/j.patrec.2010.03.014.

GHIASSI, M.; ZIMBRA, D. K.; SAIDANE, H. Urban Water Demand Forecasting with a Dynamic Artificial Neural Network Model. **Journal of Water Resources Planning and Management**, v. 134, n. 2, p. 138–146, mar. 2008.

GOODCHILD, C. W. Modelling the impact of climate change on domestic water demand. **Water and Environment Journal**, [s.l.], v. 17, n. 1, p. 8-12, mar. 2003. Wiley. http://dx.doi.org/10.1111/j.1747-6593.2003.tb00423.x.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of machine learning research**, v. 3, n. Mar, p. 1157-1182, 2003.

HALL, M. A. **Correlation-based Feature Selection for Machine Learning.** 1999. 198 f. Tese (Doutorado) - Department of Computer Science, University of Waikato, Hamilton, 1999.

HANSEN, C. W.; STRULIK, H. Life expectancy and education: evidence from the cardiovascular revolution. **Journal of Economic Growth**, [s.l.], v. 22, n. 4, p. 421-450, 4 set. 2017. Springer Nature America, Inc. http://dx.doi.org/10.1007/s10887-017-9147-x.

HERRERA, M. et al. Predictive models for forecasting hourly urban water demand. **Journal of Hydrology**, v. 387, n. 1-2, p. 141-150, 2010.

HOUSE-PETERS, L. A.; CHANG, H. Urban water demand modeling: Review of concepts, methods, and organizing principles. **Water Resources Research**, v. 47, n. 5, p. 351–360, maio 2011.

HOUSE-PETERS, L.; PRATT, B.; CHANG, H. Effects of Urban Spatial Structure, Sociodemographics, and Climate on Residential Water Consumption in Hillsboro, Oregon1. **Jawra Journal of the American Water Resources Association**, [s.l.], v. 46, n. 3, p. 461-472, 29 jan. 2010. Wiley. http://dx.doi.org/10.1111/j.1752-1688.2009.00415.x.

JAIN, A.; VARSHNEY, A. K.; JOSHI, U. C. Short-Term Water Demand Forecast Modelling at IIT Kanpur Using Artificial Neural Networks. **Water Resources Management**, [s.l.], v. 15, n. 5, p. 299-321, 2001. Springer Nature. http://dx.doi.org/10.1023/a:1014415503476.

JOHN, G. H.; KOHAVI, R.; PFLEGER, K. Irrelevant Features and the Subset Selection Problem. **Machine Learning Proceedings 1994**, [s.l.], p. 121-129, 1994. Elsevier. http://dx.doi.org/10.1016/b978-1-55860-335-6.50023-4.

KARSOLIYA, S. Approximating number of hidden layer neurons in multiple hidden layer BPNN architecture. **International Journal of Engineering Trends and Technology**, v. 3, n. 6, p. 714-717, 2012.

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. **Artificial Intelligence**, [s.l.], v. 97, n. 1-2, p. 273-324, dez. 1997. Elsevier BV. http://dx.doi.org/10.1016/s0004-3702(97)00043-x.

KOO, J. et al. Estimating regional water demand in Seoul, South Korea, using principal component and cluster analysis. **Water Science and Technology**: Water Supply, [s.l.], v. 5, n. 1, p. 1-7, mar. 2005. IWA Publishing. http://dx.doi.org/10.2166/ws.2005.0001.

KVALHEIM, O. M. Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots. **Journal of Chemometrics**, [s.l.], v. 24, n. 7-8, p. 496-504, 17 fev. 2010. Wiley. http://dx.doi.org/10.1002/cem.1289.

LEE, S.; WENTZ, E. A. Applying Bayesian Maximum Entropy to extrapolating local-scale water consumption in Maricopa County, Arizona. **Water Resources Research**, [s.l.], v. 44, n. 1, p. 1-13, jan. 2008. American Geophysical Union (AGU). http://dx.doi.org/10.1029/2007wr006101.

LI, H. et al. Recipe for revealing informative metabolites based on model population analysis. **Metabolomics**, [s.l.], v. 6, n. 3, p. 353-361, 1 maio 2010. Springer Nature. http://dx.doi.org/10.1007/s11306-010-0213-z.

LI, W.; HUICHENG, Z. Urban water demand forecasting based on HP filter and fuzzy neural network. **Journal of Hydroinformatics**, [s.l.], v. 12, n. 2, p.172-184, mar. 2010. IWA Publishing. http://dx.doi.org/10.2166/hydro.2009.082.

LIAW, A.; WIENER, M. Classification and Regression by randomForest. **Rnews**, [s.l.], v. 2/3, p. 18-22, dez. 2002.

LIPPMANN, R. An introduction to computing with neural nets. **IEEE ASSP Magazine**, [s.l.], v. 4, n. 2, p. 4-22, 1987. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/massp.1987.1165576.

LIU, H.; YU, L. Toward integrating feature selection algorithms for classification and clustering. **IEEE Transactions on Knowledge and Data Engineering**, [s.l.], v. 17, n. 4, p. 491-502, abr. 2005. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/tkde.2005.66.

MAIDMENT, D. R.; MIAOU, S. Daily Water Use in Nine Cities. **Water Resources Research**, [s.l.], v. 22, n. 6, p. 845-851, jun. 1986. American Geophysical Union (AGU). http://dx.doi.org/10.1029/wr022i006p00845.

MEHMOOD, T. et al. A Partial Least Squares based algorithm for parsimonious variable selection. **Algorithms for Molecular Biology**, [s.l.], v. 6, n. 1, p. 1-12, dez. 2011. Springer Nature. http://dx.doi.org/10.1186/1748-7188-6-27.

MEHMOOD, T. et al. A review of variable selection methods in Partial Least Squares Regression. **Chemometrics and Intelligent Laboratory Systems**, [s.l.], v. 118, p. 62-69, ago. 2012. Elsevier BV. http://dx.doi.org/10.1016/j.chemolab.2012.07.010.

MIAOU, S. A class of time series urban water demand models with nonlinear climatic effects. **Water Resources Research**, [s.l.], v. 26, n. 2, p. 169-178, fev. 1990. American Geophysical Union (AGU). http://dx.doi.org/10.1029/wr026i002p00169.

MOHAMED, M. M.; AL-MUALLA, A. A. Water Demand Forecasting in Umm Al-Quwain (UAE) Using the IWR-MAIN Specify Forecasting Model. **Water Resources Management**, [s.l.], v. 24, n. 14, p. 4093-4120, 6 maio 2010. Springer Nature. http://dx.doi.org/10.1007/s11269-010-9649-1.

ÖZESMI, S. L; ÖZESMI, U. An artificial neural network approach to spatial habitat modelling with interspecific interaction. **Ecological Modelling**, [s.l.], v. 116, n. 1, p. 15-31, mar. 1999. Elsevier BV. http://dx.doi.org/10.1016/s0304-3800(98)00149-5.

PNUD; IPEA; FJP. Atlas do desenvolvimento humano nas regiões metropolitanas. Brasília: PNUD, 2014. Disponível em: <http://atlasbrasil.org.br/2013/data/rawData/publicacao_atlas_rm_en.pdf > Acesso em: 13 jan. 2019.

POLEBITSKI, A. S.; PALMER, R. N. Seasonal Residential Water Demand Forecasting for Census Tracts. **Journal of Water Resources Planning and Management**, [s.l.], v. 136, n. 1, p. 27-36, jan. 2010. American Society of Civil Engineers (ASCE). http://dx.doi.org/10.1061/(asce)wr.1943-5452.0000003.

PREFEITURA MUNICIPAL DE FORTALEZA. **Plano Fortaleza 2040.** Fortaleza: Iplanfor, 2016.

PULIDO-CALVO, I. et al. Linear regressions and neural approaches to water demand forecasting in irrigation districts with telemetry systems. **Biosystems Engineering**, [s.l.], v. 97, n. 2, p. 283-293, jun. 2007. Elsevier BV. http://dx.doi.org/10.1016/j.biosystemseng.2007.03.003.

RAJALAHTI, T. et al. Discriminating Variable Test and Selectivity Ratio Plot: Quantitative Tools for Interpretation and Variable (Biomarker) Selection in Complex Spectral or Chromatographic Profiles. **Analytical Chemistry**, [s.l.], v. 81, n. 7, p. 2581-2590, abr. 2009. American Chemical Society (ACS). http://dx.doi.org/10.1021/ac802514y.

REED, R.; MARKSII, R. J. **Neural smithing: supervised learning in feedforward artificial neural networks**. Mit Press, 1999.

ROSENBERG, D. E. et al. Modeling integrated water user decisions in intermittent supply systems. **Water Resources Research**, [s.l.], v. 43, n. 7, p. 1-15, jul. 2007. American Geophysical Union (AGU). http://dx.doi.org/10.1029/2006wr005340.

ROSIPAL, R.; KRÄMER, N. Overview and recent advances in partial least squares. In: **International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"**. Springer, Berlin, Heidelberg, 2005. p. 34-51.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **Nature**, [s.l.], v. 323, n. 6088, p. 533-536, out. 1986. Springer Nature. http://dx.doi.org/10.1038/323533a0.

SCHLEICH, J.; HILLENBRAND, T. Determinants of residential water demand in Germany. **Ecological Economics**, [s.l.], v. 68, n. 6, p. 1756-1769, abr. 2009. Elsevier BV. http://dx.doi.org/10.1016/j.ecolecon.2008.11.012.

SHANDAS, V.; PARANDVASH, G. H. Integrating Urban Form and Demographics in Water-Demand Management: An Empirical Case Study of Portland, Oregon. **Environment and Planning B**: Planning and Design, [s.l.], v. 37, n. 1, p. 112-128, fev. 2010. SAGE Publications. http://dx.doi.org/10.1068/b35036.

TRAN, T. N. et al. Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC). **Chemometrics and Intelligent Laboratory Systems**, [s.l.], v. 138, p. 153-160, nov. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.chemolab.2014.08.005.

WANG, X. et al. An eco-environmental water demand based model for optimising water resources using hybrid genetic simulated annealing algorithms. Part II. Model application and results. **Journal of Environmental Management**, [s.l.], v. 90, n. 8, p. 2612-2619, jun. 2009. Elsevier BV. http://dx.doi.org/10.1016/j.jenvman.2009.02.009.

WENTZ, E. A.; GOBER, P. Determinants of Small-Area Water Consumption for the City of Phoenix, Arizona. **Water Resources Management**, [s.l.], v. 21, n. 11, p. 1849-1863, 2 fev. 2007. Springer Nature. http://dx.doi.org/10.1007/s11269-006-9133-0.

WOLD, S. JOHANSSON, E. COCCHI, M. in: PLS: **Partial Least Squares Projections to Latent Structures**, 3D QSAR in drug design, 1993, p. 523–550.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. **Chemometrics And Intelligent Laboratory Systems**, [s.l.], v. 58, n. 2, p. 109-130, out. 2001. Elsevier BV. http://dx.doi.org/10.1016/s0169-7439(01)00155-1.

ZHOU, S. I. et al. Forecasting daily urban water demand: a case study of Melbourne. **Journal of Hydrology**, [s.l.], v. 236, n. 3-4, p. 153-164, set. 2000. Elsevier BV. http://dx.doi.org/10.1016/s0022-1694(00)00287-0.