FEDERAL UNIVERSITY OF CEARÁ

CENTER OF TECHNOLOGY

DEPARTMENT OF HYDRAULICS AND ENVIRONMENTAL ENGINEERING

GRADUATE PROGRAM IN CIVIL ENGINEERING: WATER RESOURCES

DOCTORAL DEGREE IN CIVIL ENGINEERING: WATER RESOURCES

TAÍS MARIA NUNES CARVALHO

MACHINE LEARNING FOR WATER RESOURCES MANAGEMENT

FORTALEZA

2023

TAÍS MARIA NUNES CARVALHO

MACHINE LEARNING FOR WATER RESOURCES MANAGEMENT

Thesis submitted to the Graduate Program in Civil Engineering: Water Resources of the Center of Technology of the Federal University of Ceará, as a partial requirement for obtaining the title of Doctor in Civil Engineering. Concentration Area: Water Resources

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

FORTALEZA

2023

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

N929m Nunes Carvalho, Taís Maria.
Machine Learning for Water Resources Management / Taís Maria Nunes Carvalho. – 2023.
267 f. : il. color.

Tese (doutorado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia Civil: Recursos Hídricos, Fortaleza, 2023.
Orientação: Prof. Dr. Francisco de Assis de Souza Filho.

1. Water resources management. 2. Water demand. 3. Water quality. 4. Machine learning. 5. Statistical learning. I. Título.

CDD 627

TAÍS MARIA NUNES CARVALHO

MACHINE LEARNING FOR WATER RESOURCES MANAGEMENT

Thesis submitted to the Graduate Program in Civil Engineering: Water Resources of the Center of Technology of the Federal University of Ceará, as a partial requirement for obtaining the title of Doctor in Civil Engineering. Concentration Area: Water Resources

Approved on:

EXAMINATION BOARD

Prof. Dr. Francisco de Assis de Souza Filho (Advisor)
Federal University of Ceará - UFC

Prof. Dr. Iran Eduardo Lima Neto
Federal University of Ceará - UFC

Profa. Dra. Ticiana Marinho de Carvalho Studart
Federal University of Ceará - UFC

Dra. Mariana Madruga de Brito
Helmholtz Centre for Environmental Research

Prof. Dr. Dirceu Silveira Reis Junior
University of Brasília - UnB

To my parents Terezinha and Cláudio. To my grandfather Antônio Gomes da Silva (in memorian).

ACKNOWLEDGEMENTS

Many people walked together with me while I was constructing this thesis, and for this, I will always be grateful.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 88882.344015/2019-01. I thank CAPES for the scholarship I was awarded to conduct this thesis. I would never be able to do a PhD without financial support, and I feel privileged to have had this opportunity. During my PhD, we faced one of the most unstable political moments of the history. Universities were constantly being attacked and the credibility of researchers was severely questioned by many people. Many times, I doubted if I was doing right by pursuing a career in science. However, I was always very sure that education was the only way to fight misinformation and a path to create space for more people to have a say in the future. The student support programs, scholarships, extension projects and language courses provided by the Federal University of Ceará were essential for me to be finishing a PhD.

I thank the Post-Graduation Program in Civil Engineering and all its professors and staff. Getting accepted to this program was a dream come true. I enjoyed every class and every time I had the chance to be in one of the classrooms learning from people I admire so much. I thank all the members of the examination committee, Prof. Ticiana Studart, Prof. Iran Lima Neto, Prof. Dirceu Reis and Dr. Mariana de Brito, for accepting to examine this thesis. I sincerely admire all of them. I thank Prof. Samiria for her support and guidance in all the research projects we worked together. I am grateful for the advices and shared knowledge during lunch breaks. I thank Prof. Ticiana for her guidance and collaboration in every shared project and papers we have worked together. I thank Prof. Iran for his kindness and careful guidance in one of the papers of this thesis.

I thank my mom, Terezinha, for supporting me in every moment of this path and for being my best friend. She inspires me in every possible aspect and I could never finish this PhD without her. Her kindness, patience and unconditional love give me reason to keep studying and working everyday. I also thank my dad, Cláudio, for always cheering for me and showing the best side of things when I could not see it myself. I am grateful for the beautiful trees and flowers that he so carefully cultivates in our house and that were inspirational for me so many times. His love and care give me reason to keep going.

I thank my brother, João Cláudio, for supporting my education in every possible way.

I thank my sister, Taiana Cláudia, for filling my life with music and nature and reminding myself to enjoy small moments. I am happy to have them by my side looking out for me.

I thank my aunts Íris and Alvanisia for their love and support.

I thank my grandfather Antônio Gomes, who will always inspire me with his strength and intelligence. His history as a workman building some of Cearás's dams makes me proud and happy to be working in the water resources sector. I thank my grandmother Maria Nunes for her strength and serenity.

I thank my partner in life Walter, for the unlimited support and for always reminding myself what I can achieve. I am grateful for walking this academic journey with him and being inspired by him in so many steps of the way.

I thank Cida for her friendship and companionship. I am so grateful to have her as my friend since day one at UFC. I thank my friends and lab colleagues Ályson and Tereza, for the friendship and for sharing this academic path with me. It is always inspiring talking to you about research ideas and hearing your inputs. I thank my friend Priscila Natiele and her family for their kindness and for being like family to me. I thank my childhood friends Renata Freire and Letícia Sampaio. Their friendship mean a lot to me and sure made this path much more happy.

Lastly, I thank my advisor Prof. Francisco de Assis for mentoring me since my undergraduate studies. You inspired me to follow an academic career with your distinct love for teaching and research. I thank you for the guidance, patience and devotion during this journey. Thank you for going beyond technical education and bringing literature, poetry and national music to our meetings. We might still live as our parents, but you helped me to fill my toolbox with tools to fight the dangers around the corner. It has been a honour learning from you.

"De cada vivimento que eu real tive, de alegria forte ou pesar, cada vez daquela hoje vejo que eu era como se fosse diferente pessoa. Sucedido desgovernado. Assim eu acho, assim é que eu conto."

(João Guimarães Rosa)

ABSTRACT

Water resources management challenges are multiple and complex, and human force, as the main driver of environmental change, has been increasing the need for tailored (and faster) responses to them over the past decades. Despite our increasing technical knowledge on how to tackle these issues, which are mainly related to water quantity, quality and access, unprecedented change in climate and landscapes will require a better understanding of the interactions between water and society. This thesis is concerned with the challenging task of applying machine learning techniques to extract knowledge from hydrological, socioeconomic and climate data and tackle some of the water management issues associated with water quantity and quality. Specifically, it addresses (i) the drivers of water demand in different temporal and spatial scales; (ii) the implications of price-based demand-side measures and how media coverage and public interest on extreme events can affect consumption habits; (iii) the long-term water availability and supply under climate variability, and (iv) some of the effects of environmental change on water quality. We learn that climate variability and change might affect not only hydrological responses but also consumption habits and water supply expansion strategies. Also, we make valuable findings on the drivers of water demand and quality, which can support utilities in their long-term planning.

Keywords: Water resources management. Water demand. Water quality. Machine learning. Statistical learning.

RESUMO

Os desafios da gestão de recursos hídricos são múltiplos e complexos, e a força humana, como principal impulsionadora da mudança ambiental, tem aumentado a necessidade de respostas personalizadas (e mais rápidas) para eles nas últimas décadas. Apesar de nosso crescente conhecimento técnico sobre como lidar com essas questões, principalmente relacionadas à quantidade, qualidade e acesso à água, mudanças sem precedentes no clima e no uso da terra exigirão uma melhor compreensão das interações entre a água e a sociedade. Esta tese está preocupada com a tarefa desafiadora de aplicar técnicas de aprendizado de máquina para extrair conhecimento de dados hidrológicos, socioeconômicos e climáticos e abordar alguns dos problemas de gerenciamento de água associados à quantidade e qualidade da água. Especificamente, aborda (i) as variáveis que influenciam a demanda de água em diferentes escalas temporais e espaciais; (ii) as implicações de medidas de controle da demanda baseadas em preços e como a cobertura da mídia e o interesse público em eventos extremos podem afetar os hábitos de consumo de água; (iii) a disponibilidade e abastecimento de água a longo prazo sob variabilidade climática, e (iv) alguns dos efeitos da mudança ambiental na qualidade da água. Aprendemos que a variabilidade e as mudanças climáticas podem afetar não apenas as respostas hidrológicas, mas também os hábitos de consumo e as estratégias de expansão do abastecimento de água. Além disso, fazemos descobertas valiosas sobre os impulsionadores da demanda e qualidade da água, que podem apoiar as concessionárias de água em seu planejamento de longo prazo.

Palavras-chave: Gestão dos recursos hídricos. Demanda hídrica. Qualidade da água. Aprendizado de máquina. Aprendizado estatístico.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

R^2	R-squared
ADF	Augmented Dickey-Fuller
ALE	accumulated local effect
ANA	Brazilian Water and Sanitation Agency
ANN	Artificial neural network
ARMA	autoregressive and moving average
CAGECE	Ceará Water and Wastewater Company
CB	census block
Chla	Chlorophyll-a concentrations
COGERH	Water Resources Management Company of Ceará
CRU	Climatic Research Unit
CT	census tract
ECDF	Empirical Cumulated Distribution Function
EEMD	ensemble empirical mode decomposition
EMD	empirical mode decomposition
GBM	Gradient boosting machines
GBR	gradient boosting regression
GLM	generalized linear model
IBGE	Brazilian Institute of Geography and Statistics
IBR	increasing block rates
IIS	iterative input selection
IMF	intrinsic mode function
IMFs	intrinsic mode functions
IncMSE	increase in the MSE
IRD	*Research Institute for Development* / Institut de Recherche pour le Développement
IWSS	Integrated Water Supply System
JMS	Jaguaribe-Metropolitano supply system
kNN	k-Nearest Neighbors
LOESS	locally weighted smoothing
MAE	Mean Absolute Error
MHDI	Human Development Index

MHDI	Municipal HDI
MI	mutual information
MISO	multi-input single-output model
ML	Machine learning
MLP	multilayer perceptron network
MSE	Mean Squared Error
PACF	partial autocorrelation function
PDP	partial dependence plot
PISF	*São Francisco River Transposition Project* / Projeto Integrado de Transposição do Rio São Francisco
RF	random forest
RMF	*Região Metropolitana de Fortaleza* / Metropolitan Region of Fortaleza
RMSE	Root Mean Square Error
SDDP	Stochastic Dual Dynamic Programming
SISO	single-input-single-output
SOM	Self-organizing map
SSR	squares of the residuals
STL	Locally estimated scatterplot smoothing
VMD	variational mode decomposition

CONTENTS

1 INTRODUCTION

Water resources management challenges are multiple and complex, and human force, as the main driver of environmental change (COSGROVE; LOUCKS, 2015), has been increasing the need for tailored (and faster) responses to them over the past decades. Despite our increasing technical knowledge on how to tackle these issues, which are mainly related to water quantity, quality and access (LOUCKS *et al.*, 2017), unprecedented change in climate and landscapes will require a better understanding of the interactions between water and society.

At the same time, this intense social and environmental transformation has been accompanied by an exponential growth of computational resources and data, be it human- or machine-generated (SIT *et al.*, 2020). In fact, in the last decades, new data collection strategies have been adopted by the water sector, such as satellite remote sensing (MUSA *et al.*, 2015), "smart" meters (COMINOLA *et al.*, 2015), crowdsourcing approaches (WEESER *et al.*, 2018) and text mining (BRITO *et al.*, 2020). Statistical learning comprise a set of tools to process this data and gain insight from it (HASTIE *et al.*, 2009). This means we might able to learn from the past, not by simply expecting the future to be the same, but by assessing how nature might respond to our actions and vice versa.

Machine learning (ML) techniques have been intensely explored in the water resources field, especially for predicting hydrological and hydroclimatological variables, such as streamflow (PAPACHARALAMPOUS; TYRALIS, 2018; LIN *et al.*, 2021; XU *et al.*, 2022), precipitation (WEI *et al.*, 2022; TAO *et al.*, 2021), and water demand (BRENTAN *et al.*, 2017; GHARABAGHI *et al.*, 2019; DUERR *et al.*, 2018); for unraveling hydrological processes (SCHäFER *et al.*, 2022); improving conceptual hydrological modeling (KUMANLIOGLU; FISTIKOGLU, 2019), and quantifying hydrological extremes (HAUSWIRTH *et al.*, 2021). However, less effort has been put into incorporating these models into water resources management and planning tasks.

One reason for that is the black-box nature of most ML models: their either have incomprehensible underlying functions, or are constructed with abstract features, as in the case of deep learning. While these models might offer accurate predictions of hydrological and environmental variables, their supposed lack of interpretability makes them less suitable - and thus less explored - to make decisions (RUDIN, 2019). They can, however, be used to guide the learning process and the development of water management modes. In this context, there are additional barriers to integrate them into decision making: the theoretical basis of these models

is not part of the traditional background of most water resources practitioners and this might affect their credibility with stakeholders (SHEN, 2018). In fact, decisions of water management institutions have been mainly guided by process-based models so that incorporating purely data-driven models into it should be challenging (OLSSON; ANDERSSON, 2007). We argue that if combined with expert knowledge and used with parsimony, ML can be a useful tool to advance science for planning and managing water resources systems, since the complex dynamic of the hydrological and social systems (FICKLIN *et al.*, 2022) require modeling approaches capable of dealing with nonlinearities (KUMAR, 2015). We make a first attempt to show how ML can be leveraged to guide water management and the learning process of water resources stakeholders.

This thesis is concerned with the task of applying ML techniques to extract knowledge from hydrological and socioeconomic data and tackle some of the water management issues associated with water quantity and quality. Specifically, it addresses (1) the drivers of water demand in different temporal and spatial scales; (2) the implications of price-based demand-side measures and how media coverage and public interest on extreme events can affect consumption habits; (3) strategies to manage water availability and supply under climate variability in the long-term, and (4) some of the effects of environmental change on water quality.

We combine several ML algorithms to explore the relationships between social, economic, climatological and hydrological attributes and variables of interest to water resources management. Beyond obtaining accurate predictions, we use different tools to improve interpretability of machine learning and rule extraction to gain significant insights on the human-water interfaces. Although most of the ML algorithms used in this thesis are already widely known, we show how water resources managers and stakeholders can benefit from them and reduce the impacts of climate variability on water security. We present innovative applications and strategies to take advantage of state of the art tools that have not yet been sufficiently explored in a way that can be easily replicated for different social and environmental contexts. We show that ML can help researchers to explore characteristics of water users, quantify the effects of water policies, decide which pathways to choose when developing a plan and verify the role of climate and hydrology on eutrophication.

1.1 Roadmap

Chapter 2 presents the methodological roadmap for all studies conducted in this thesis, which are mainly based on statistical learning theory. Chapter 3 presents the sources and spatial levels of the data used here. Chapters 4 and 5 refer to objective 1, Chapter 6 to objective 2, Chapter 7 to objective 3, and Chapter 8 to objective 4. All data is available in a public dashboard.

- Chapter 4 is based on the study entitled "Urban Water Demand Modeling Using Machine Learning Techniques". In this chapter, machine learning methods are combined to explore the main drivers of water demand. The entire code is available on GitHub.
- Chapter 5 is based on the study entitled "Variational mode decomposition hybridized with gradient boost regression for seasonal forecast of residential water demand". In this chapter, we present an original method for seasonal forecast of water demand.
- Chapter 6 is based on the study entitled "A Data-Driven Model to Evaluate the Medium-Term Effect of Contingent Pricing Policies on Residential Water Demand".
- Chapter 7 is based on the work developed for the project "Optimization of the water supply system of Fortaleza and inclusion of alternative water sources".
- Chapter 8 is based on the study entitled "Uncovering the Influence of Hydrological and Climate Variables in Chlorophyll-A Concentration in Tropical Reservoirs with Machine Learning". The entire code is available on GitHub.

Chapter 9 summarizes the main conclusions of this thesis and highlights the outcomes and implications of the conducted research.

2 METHODOLOGICAL ROADMAP

Statistical learning tools are designed to learn from data and make predictions. Depending on the data available and the problem one is trying to solve, these tools can be generally classified as (i) supervised learning, (ii) unsupervised learning or (iii) reinforcement learning. When learning occurs from a training set of data, where each occurrence has its own correspondent label, we might have a supervised learning problem. Learning means finding the function that better maps the input and output (label).

Supervised learning tasks can be formulated as either regression or classification problems, depending on the output. If the output takes a continuous range of values, it is a regression problem. Using water demand prediction as an example (Chapter 4), a regression could be performed with average per capita income as input and water demand as an output. The regression would find a functional relationship between average per capita income (I) and water demand (D), such that:

$$D = f(I)$$

More than one variable might be necessary to explain the outcome, and in this case, the function becomes more complex and more data might be needed to train the model. In the problems described in Chapters 4 and 8, for instance, water demand and Chlorophyll-a concentrations were modeled from several socioeconomic and hydroclimatological variables, respectively. If the task involves dealing with sequential data, regression models can still be useful for providing forecasts, as long as data points are not considered to be independent (Chapter 5 and Chapter 6).

In classification problems, the output is extracted from a discrete set of labels. In water resources problems, for example, a set of water quality parameters of a lake would be the input, and the output label would be that lake's water quality index. Some algorithms (e.g. classification and regression tree and artificial neural networks) allow us to extract rules that explicitly describe the relationships between input and output variables (Chapter 8). Unsupervised learning is appropriate when the data has no label or a specific desired output. In Chapter 4, for example, we use a clustering algorithm to identify water consumption profiles in a urban environment.

Table 1 summarizes the models and algorithms used in this thesis and in which chapters each of them was applied. All models (except for the optimization model in chapter 7) were developed using R programming language. The packages used to perform data manipulation

and preparation (e.g. dplyr and tidyr), modeling and visualization (e.g. ggplot2 and rpart.plot) are summarized in Table 2.

Table 1 – Machine Learning models and algorithms used in this thesis.

Model	Section with the corresponding description	Chapter				
		4	5	6	7	8
Random Forest	4.3.2, 8.2.8	✓	×	×	×	✓
Self-organizing Map	4.3.4	✓	×	×	×	×
Artificial Neural Network	4.3.6, 8.2.5	✓	×	×	×	✓
Accumulated Local Effect	4.3.3	✓	×	×	×	×
Variational Mode Decomposition	5.3.1	×	✓	×	×	×
Gradient Boosting Machine	5.3.2, 5.3.2, 8.2.8	×	✓	✓	×	✓
Partial Autocorrelation Function	5.3.3	×	✓	×	×	×
Partial Dependence	6.3.4, 8.5	×	×	✓	×	✓
Seasonal-Trend decomposition using LOESS	8.2.9	×	×	✓	×	×
Support Vector Machine	8.2.9	×	×	×	×	✓
Linear Regression	8.2.3	×	×	×	×	✓
Elastic-Net Regularized Generalized Linear Model	8.2.4	×	×	×	×	✓
kNN	8.2.6	×	×	×	×	✓
Classification and Regression Tree	8.2.7	×	×	×	✓	✓

Source: The author.

Table 2 – R packages used in this thesis.

R Package	Reference	Chapter				
		4	5	6	7	8
dplyr	Wickham *et al.* (2022)	✓	✓	✓	✓	✓
tidyr	Wickham (2020)	✓	✓	✓	✓	✓
purrr	Henry and Wickham (2020)	✓	✓	✓	✓	✓
ggplot2	Wickham (2016)	✓	✓	✓	✓	✓
magrittr	Bache and Wickham (2020)	✓	✓	✓	✓	✓
corrplot	Wei and Simko (2017)	✓	×	×	×	✓
randomForest	Liaw and Wiener (2002)	✓	×	×	×	✓
kohonen	Wehrens and Kruisselbrink (2018)	✓	×	×	×	×
RSNNS	Bergmeir and Benítez (2012)	✓	×	×	×	✓
ALEPlot	Apley (2018)	✓	×	×	×	×
vmd	Hamilton and Ferry (2017)	×	✓	×	×	×
gbm	Greenwell *et al.* (2020)	×	✓	✓	×	✓
pdp	6.3.4, Greenwell (2017)	×	×	✓	×	✓
e1071	Meyer *et al.* (2020)	×	×	×	×	✓
kNN	8.2.6	×	×	×	×	✓
rpart	Therneau and Atkinson (2019)	×	×	×	×	✓
rpart.plot	Milborrow (2020)	×	×	×	×	✓

Source: The author.

3 DATA

Several databases were used to train the models described here (Table 4 and 3). Social and economic data were obtained mainly from the Brazilian Institute of Geography and Statistics (IBGE) census, but also from Fortaleza's planning Institute in geospatial vector data format. Hydrological data (e.g. precipitation and temperature) were obtained from different sources, including (i) climate stations, (ii) reanalysis databases, and (iii) measured data. Water quality information was extracted from satellite-based data and field measures.

Table 3 – Variables, sources and data types obtained in the NetCDF format used in this thesis.

Variable	Spatial level	Time period	Source
Mean surface temperature over the reservoir	0.5 degree grid	2002-2019	CRU Harris *et al.* (2020)
Monthly average of surface and subsurface runoff accumulated over one day in the hydrographic basin	9 km grid	2002-2019	Muñoz-Sabater *et al.* (2021)
Air temperature at 2 m above the reservoir	9 km grid	2002-2019	Muñoz-Sabater *et al.* (2021)
Water temperature at the bottom of the reservoir	9 km grid	2002-2019	Muñoz-Sabater *et al.* (2021)
Thickness of the mixed layer	9 km grid	2002-2019	Muñoz-Sabater *et al.* (2021)
Surface net solar radiation	9 km grid	2002-2019	Muñoz-Sabater *et al.* (2021)
Horizontal wind speed at a height of 10 m above the reservoir surface	9 km grid	2002-2019	Muñoz-Sabater *et al.* (2021)

Source: The author.

Table 4 – Variables, sources and data types obtained from tabular databases used in this thesis.

Variable	Spatial level	Time period	Source
Census data			
Human Development Index	Neighborhood	2010	IPLANFOR
Average per capita income	Census tract	2010	IBGE
% Female residents	Census block	2010	PNUD (2012)
% 65 years old or older	Census block	2010	IBGE
% 1 to 14 years old	Census block	2010	PNUD (2012)
Life expectancy	Census block	2010	PNUD (2012)
Expected years of schooling	Census block	2010	PNUD (2012)
% 25 years or older who have completed Elementary School	Census block	2010	PNUD (2012)
% 25 years or older who have completed High School	Census block	2010	PNUD (2012)
% 25 years or older who have completed College	Census block	2010	PNUD (2012)
Average per capita income	Census block	2010	PNUD (2012)
% Population living in poverty	Census block	2010	PNUD (2012)
% Population vulnerable to poverty	Census block	2010	PNUD (2012)
% Population living in households with bathrooms and running water	Census block	2010	PNUD (2012)
% Population living in urban households with a garbage collection service	Census block	2010	PNUD (2012)
% People in households with inadequate water supply and sanitation facilities	Census block	2010	PNUD (2012)
% Economically active population aged 18 or older	Census block	2010	PNUD (2012)
% People in households vulnerable to poverty in which no one has completed Elementary School	Census block	2010	PNUD (2012)
Municipal Human Development Index	Census block	2010	PNUD (2012)
Demographic density	Census tract	2019	IPLANFOR
Measured data			
Average monthly precipitation	Point	2010	HidroWeb (2010)
Water level	Reservoir	2002-2019	COGERH
Total water volume in the reservoir	Reservoir	2002-2019	COGERH

Source: The author.

4 URBAN WATER DEMAND MODELING USING MACHINE LEARNING TECHNIQUES

> "Sertão - sabe o senhor: sertão é onde o pensamento da gente se forma mais forte do que o poder do lugar." (ROSA, 2019)

4.1 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 for some cities in developed countries (UNESCO, 2018; SAURI, 2020). 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. Either way, 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; CHANG, 2011). Previous studies have 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 has a history of multiyear droughts and water supply crisis.

The city is supplied by multiple surface water reservoirs, which are also used for irrigation and industrial purposes. Annual precipitation is low and highly variable; hence water availability is subject to climate conditions. Aiming to expand the supply system's capacity and to reduce its climate dependence, local managers are planning to install a desalination and wastewater reuse plants. The capacity expansion plan will consist in 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). The purpose of this study is to provide a framework for water demand modeling using machine learning techniques and to explore 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, 2019). Domestic water consumption in Brazil was explored in a few previous studies (BRENTAN *et al.*, 2017; DIAS *et al.*, 2018; SANT'ANA; MAZZEGA, 2018; GARCIA *et al.*, 2019). However, they were limited to 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 square regression (NAWAZ *et al.*, 2019) and 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. ML techniques have receiving increasingly attention as researchers have come to understand that these algorithms can effectively learn information from water demand data and capture nonlinear relationships between water demand and relevant variables. In recent studies, (LEE; DERRIBLE, 2020) and (BOLORINOS *et al.*, 2020) showed that ML models outperform linear methods for prediction of 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 have also 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; COMINOLA *et al.*, 2019; BOLORINOS *et al.*, 2020).

The most popular machine learning methods in water demand studies is the Artificial neural network (ANN), that has long been used because of its excellent predictive ability (VIJAI; SIVAKUMAR, 2018). Prior research explored ANN models for predicting 15-min (GUO *et al.*, 2018), weekly (BATA *et al.*, 2020; ADAMOWSKI; KARAPATAKI, 2010) and monthly water demand (FIRAT *et al.*, 2009; ALTUNKAYNAK; NIGUSSIE, 2017), residential water end-use (BENNETT *et al.*, 2013) and irrigation demand (PULIDO-CALVO *et al.*, 2007). Other studies combined ANN with different methods to improve water demand prediction, such as seasonal autoregressive integrated moving average (BATA *et al.*, 2020) and discrete wavelet transform (ALTUNKAYNAK; NIGUSSIE, 2017).

Alternative ML techniques used to model water demand are support vector ma-

chine (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 tree and random forest (VILLARIN; RODRIGUEZ-GALIANO, 2019; BOLORINOS *et al.*, 2020). RF algorithms have been standing out in water science and hydrological applications (TYRALIS *et al.*, 2019). They have been mainly used for streamflow and water quality modeling (YAJIMA; DEROT, 2018; PAPACHARALAMPOUS; TYRALIS, 2018). A few researchers applied this method for analyzing variable importance for water demand prediction (VILLARIN; RODRIGUEZ-GALIANO, 2019; BRENTAN *et al.*, 2017) and short-term forecast (VIJAI; SIVAKUMAR, 2018; CHEN *et al.*, 2017; HERRERA *et al.*, 2010).

ML techniques are also useful for pattern recognition. Self-organizing map (SOM) – a type of neural network – has been used in several water resources applications, such as ground-water level forecasting model (HASELBECK *et al.*, 2019), water quality assessment (LI *et al.*, 2018) and analysis of land use change with satellite data (QI *et al.*, 2019). SOM was also used to analyze water consumption patterns in recent studies (BRENTAN *et al.*, 2017; PADULANO; 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.*, 2008) but are sensitive to the dataset 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 dataset is extensive.

Long-term prediction is usually related to structural, social and environmental variables, such as lot size, building density, educational level and family size (CHANG *et al.*, 2010; POLEBITSKI; PALMER, 2010). Social and structural dynamics might influence changes in water use behavior, as indicated by (GONZALES; 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; AJAMI, 2017). This discussion, however, has been mostly limited to the US and Europe.

This study seeks to provide 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 have focused on evaluating the predictive power of ML models, and so far, there has been little discussion on the individual effect of sociodemographic variables on water demand, especially in developing countries. We address this issue with the application the accumulated local effects method (APLEY; ZHU, 2016) for interpreting black box models. Domestic water demand was analyzed with cross-sectional data at two spatial levels (census tract and census block). While at the census tract level (fine scale), only two variables were available, at the census block level (coarse scale), eighteen explanatory variables were used. RF was used for ranking the variables and SOM was 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 ANN was built for both spatial levels. At the census block level, the IIS method (GALELLI; CASTELLETTI, 2013) was used to select the input variables for the predictive model.

4.2 Data

This research is a cross-sectional study that compares two spatial levels of aggregation with different data availability: census tract (CT) (n = 2952) and CB (n = 182). The dataset of the CT level included only two input variables (average per capita income and the Human Development Index; HDI), while the dataset of the CB level included eighteen variables (Table 5).

Ceará Water and Wastewater Company (CAGECE) provided a dataset with monthly water consumption over the year of 2010 for a total of 878,992 households. Data was provided with a household identifier, thus could be spatially aggregated by census tracts and census blocks. The dependent variable was average daily per capita consumption for 2010, since explanatory variables were obtained for this year. We calculated it by averaging monthly household water consumption in 2010 and dividing it by the population in the census tracts and census blocks. Average daily per capita water demand in the census tracts is presented in Figure 1.

The explanatory variables were obtained from the 2010 census, conducted by the IBGE. The 2010 census collected extensive sociodemographic information of households – grouped into census tracts – from more than five thousand municipalities in Brazil. At the census tract level, publicly available data is restricted to household composition and per capita income. Household composition was intentionally excluded from the CT dataset because this

Table 5 – Explanatory variables at the CB level.

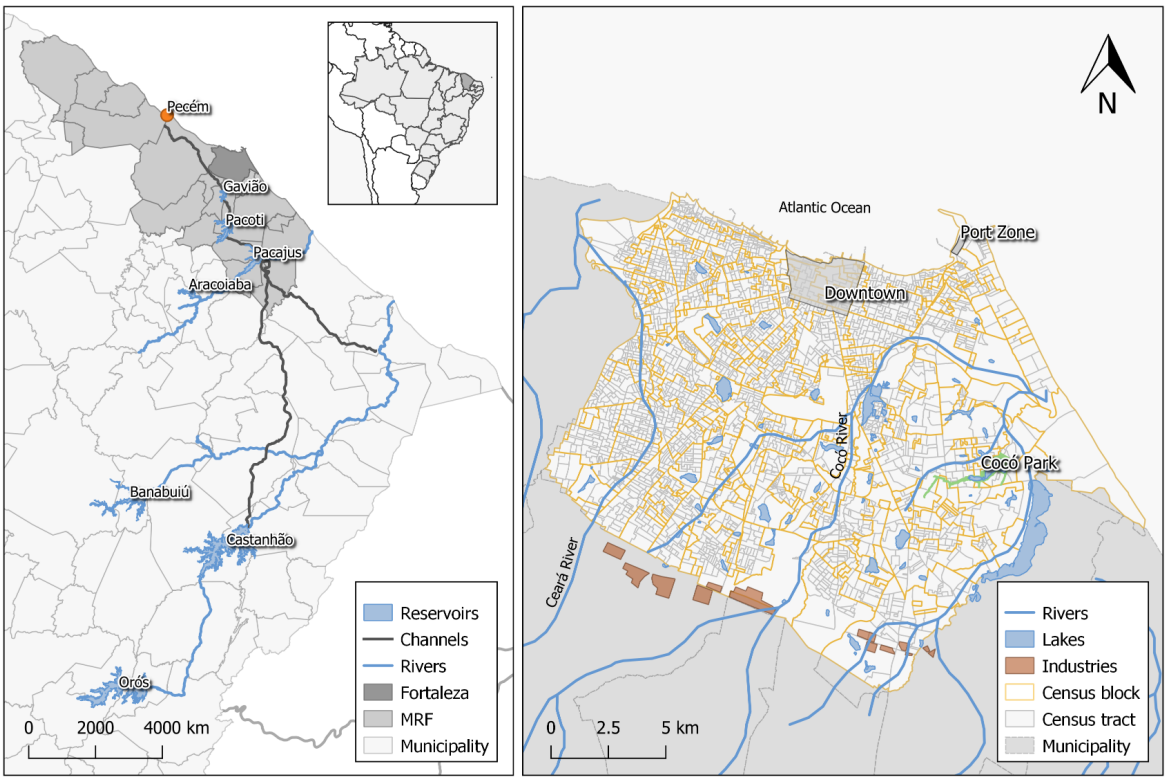
ID	Variable	Unit	Mean	St. Dev.
HDI	Human Development Index	N/D	0.362	0.194
Av. per capita income	Average per capita income	R$	2,151.15	2,424.35
Demographic variables				
% female	Female residents	%	53.32	1.82
% 65+	65 years old or older	%	6.59	2.80
% 1 to 14	1 to 14 years old	%	20.74	4.71
Dem. density	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

Source: The author.

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 alphabetized population aged 10 years or older (education) and percentage of the population over 64 years old living in the neighborhood (longevity).

Detailed census data is only released on aggregated level, for geographic units

Figure 1 – Jaguaribe-Metropolitano supply system (left) and Fortaleza's census tracts and census blocks (right).



Source: The author.

containing at least 400 households. Census blocks aggregate contiguous census tracts and are available for 23 Brazilian metropolitan areas (PNUD *et al.*, 2014). More than 200 indices are provided at this level, related to aspects of demography, education, income, employment, housing and vulnerability. Most of the indices are classified by sex and age, thus, to reduce the number of variables, some of them were merged. The final dataset included the potentially relevant variables of each category, reducing the indices to 18 variables expected (Table 5). Variables were chosen to assess socioeconomic inequalities and to explain consumer behavior. Demographic variables initially included 85 indices, narrowed down to five, 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 is perfectly correlated to percentage of women (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 corresponds to three stages: pre-school (for children from 0 to 5 years old), elementary school

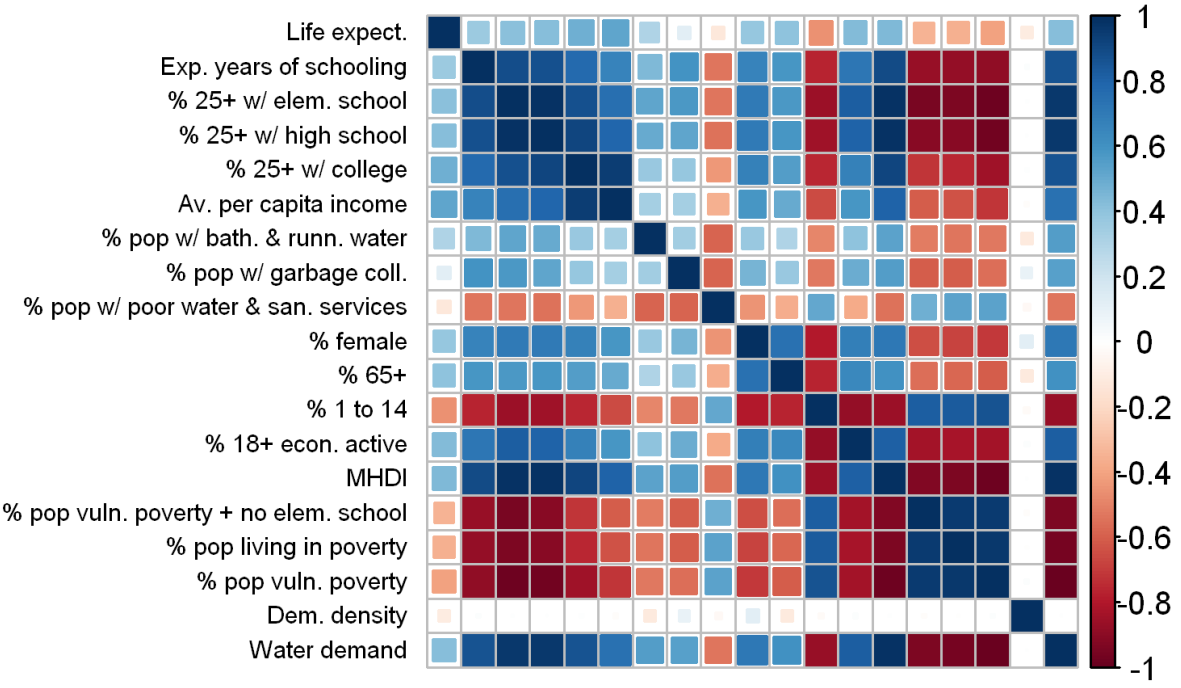
(6 to 14 years old) and high school or secondary education (15 to 17 years old). A high school diploma is mandatory for admission to higher education.

The category of income included three variables. Those considered as “living in poverty” have a per capita household income equal to or less than one-fourth of the minimum wage, while the “vulnerable to poverty” live with less than one half of the minimum wage. These variables were included because average per capita income alone could disguise information on the income gap. Variables regarding basic services for adequate housing reflect the health condition of the inhabitants (MONTGOMERY; ELIMELECH, 2007).

In the category of employment and vulnerability, the percentage of economically active population aged 18 or older accounts for the people in the job market or trying to join it. The Municipal HDI (MHDI) contemplates the same 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 profit.

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 RF (Figure 2). Except for garbage collection service, households with inadequate water supply and sanitation and demographic density, all other variables are strongly associated with water consumption. Independent variables are also correlated to each other, such as per capita income, associated with life expectancy at birth ($r = 0.74$), percentage of college educated people ($r = 0.94$) and MHDI ($r = 0.81$). Correlated variables are usually avoided because they might contain redundant information, but high correlation does not mean lack of variable complementarity (GUYON; 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 later). However, when selecting the input variables for the predictive model, the IIS method was used to avoid redundant information.

Figure 2 – Correlation matrix between independent variables (CB model) and water consumption. The square's size is proportional to the correlation between the variables.



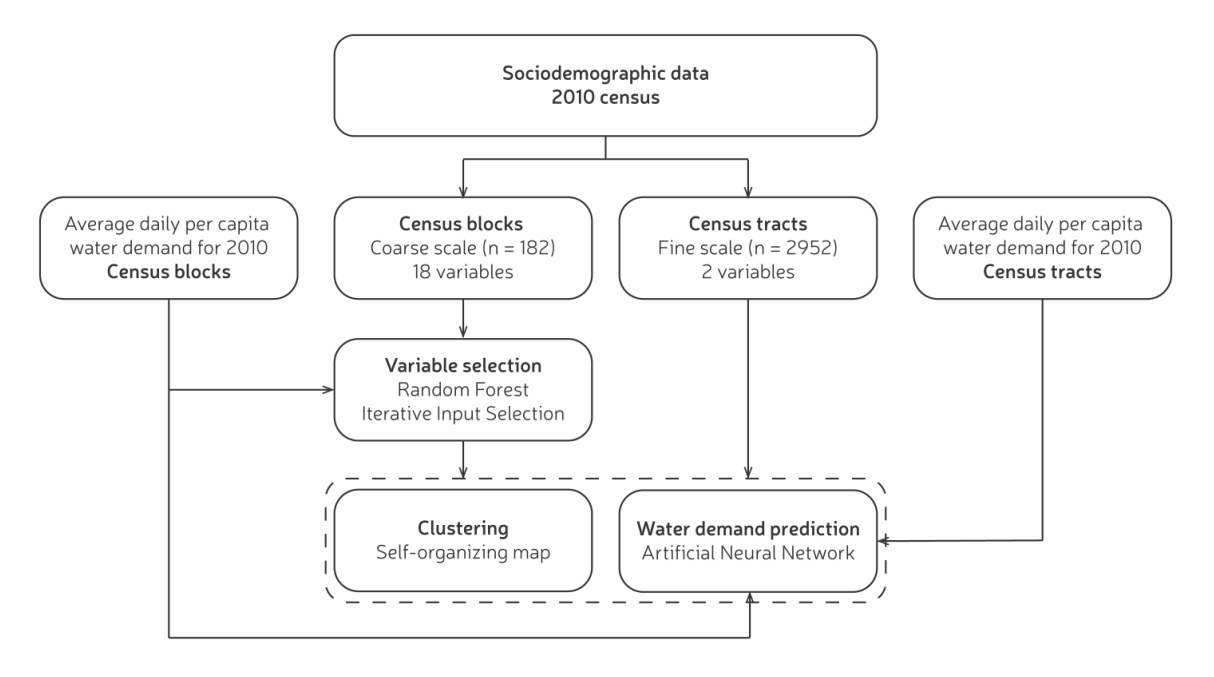
Source: The author.

4.3 Methodology

The methodology of this study is divided into three sections (Figure 3): (i) variable importance using RF; (ii) clustering and spatial analysis of demand and sociodemographic characteristics with SOM; (iii) variable selection with the IIS method and predictive model using ANN.

The first part of this study investigates which sociodemographic characteristics drive consumer behavior and water consumption. This analysis was performed at the census block level, which had 18 explanatory variables. 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 clustering was also performed with census tract data, in order to compare spatial level aggregation. The predictive model was built using an ANN and it was compared for both spatial levels: CB and CT. The first considered the variables iteratively selected with the RF and ANN models, while the last had only two explanatory variables.

Figure 3 – Methodological steps.



Source: The author.

4.3.1 Algorithms and model specifications

In this section, the machine learning models and algorithms are presented.

4.3.2 Random Forest

RF (BREIMAN, 2001) is a supervised learning algorithm mainly used for regression and classification tasks. 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 an from the original dataset. These observations will be used for building 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 ones. The number of variables to be drawn is denoted as 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, take the majority vote. For regression trees, take the average (HASTIE *et al.*, 2009):

$$y_{x_i} = \hat{f}_{RF}^{N}(x_i) = \frac{1}{N}\sum_{b=1}^{N} T_b(x_i)$$

where x_i = vector of independent variable; $T_b(x_i)$ = 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).

RF is efficient and widely used for variable selection and prediction. It is applicable to problems with nonlinear relationships between the variables and can effectively handle small sample sizes (BIAU; SCORNET, 2015). The tree-building process of random forests implicitly allows for interaction and high correlation between features (ZIEGLER; 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 MSE is computed. To assess the importance of a specific predictor variable, its values are randomly permuted for the out-of-bag sample and the MSE is computed again. The increase in the MSE (IncMSE) resulted 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 the leave-one-out cross-validation to reduce bias in training data. In this approach, one data point is left for validation and the training set is composed by 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 *et al.*, 2011); hence for this study, the error estimate was based on the average of the IncMSE for 182 tests. In order to get 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.

4.3.3 Accumulated local effect

To assess the main effects of the individual predictor variables, they were visualized with the accumulated local effect (ALE) plots (APLEY; 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,). 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 in 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,):

$$\hat{f}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i:xj(i)Njk} [f(z_{k,j}, x_{Jay}^{(i)}) - f(z_{k-1,j}, x_{Jay}^{(i)})]$$

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

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

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.

4.3.4 Self-Organizing Map

A SOM 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; 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 $w = (w_{i1}, w_{i2}, ..., w_{id})$, 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 w_i of the *mxn* neurons.
2. Randomly select an input vector $x(t)$, which represents the pattern that is presented to the neurons in the output layer.
3. Find the winner neuron c or the Best Matching Unit based on the minimum distance Euclidean criterion:

   $$c = argmin\|w_i(t) - x_t\|$$

   where $\|.\|$ is the Euclidean distance measure, x(t) and $w_i(t)$ are the input and weight vector of neuron at iteration t respectively.
4. Update the weight vector of the neurons using the following equation:

   $$w_i(t+1) = w_i(t) + h_{c,i}(t)[x(t) - w_i(t)]$$

   where $h_{c,i}(t)$ is a Gaussian neighborhood function:

   $$h_{c,i}(t) = \alpha(t) * \exp(-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)})$$

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

The main parameters of 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 have used different criteria to do it (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.

The main parameters of SOM are the grid size, the training rate and the neighborhood size. There is no theoretical justification on the literature for choosing the optimal grid size of the output layer. Previous studies have used different criteria to do it (KALTEH *et al.*, 2008), but the general recommendation is to define the size by trial-and-error. 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 6x6 network (CB level; coarse scale) and a 12x12 network (CT level; fine scale) were considered the most suitable for the problem. Larger maps resulted in many empty nodes and/or less than two data points per node.

The training rate was set to decrease linearly from 0.05 to 0.01 over 100 updates, i.e. the number of times the dataset was presented to the network. The mean radius of the neighborhood is also set to decrease linearly with the training steps. The initial neighborhood size was 3.6 points for CB and 7.0 points for CT, covering 2/3 of the distance between nodes, and the final values were zero for both models. This strategy allows SOM to be smoothed out globally, with increasing resolution (KOHONEN, 2014).

4.3.5 Cluster validation

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

$$DI = \frac{min_{1\leq i\leq m}\delta(C_i,C_j)}{max_{1\leq k\leq m}diam_{C_k}}$$

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

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

where d is the Euclidean distance. The diameter of a cluster C was defined as the Euclidean distance between the farthest two points inside the cluster. The silhouette index (ROUSSEEUW, 1987) is given by: $S_i = \frac{b_i - a_i}{max\{a_i,b_i\}}$ where a is the mean Euclidean distance between an observation and all other data points in the same cluster; and b is the mean Euclidean distance between an observation and all other points in the next cluster.

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

The clusters were also 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.

4.3.6 Artificial Neural Network

ANNs are statistical models build 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 comprised by at least two layers: input and output. The multilayer perceptron network (MLP) 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):

$$y_k = f_{outer}[\sum_{j=1}^{M} w_{kj}^{(2)} f_{inner}[\sum_{i=1}^{d} w_{ji}^{(1)} x_i + w_{j0}^{(1)}] + w_{k0}^{(2)}]$$

where y_k represents the k-th output, f_{outer} represents the output layer transfer function, f_{inner} represents the input layer transfer function, w represents the weights and biases, (i) represents the i-th layer.

The domestic water demand was projected with an MLP and trained with a back-propagation algorithm (RUMELHART *et al.*, 1986). Backpropagation 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 on how to do it (REED; MARKS, 1999), but usually, one or two hidden layers are enough to solve any nonlinear problem (LIPPMANN, 1987). An MLP with one hidden layer was used in this study. Adding more hidden layers would not only increase computational time, but also the number of parameters and a larger training dataset 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 dataset is divided into k subsets: $k-1$ are used to train the model and the remaining 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. In this study, 5 folds were used. Because variables do not commensurate, data was normalized by min-max scaling. The parameters used for performance evaluation were: Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and R^2.

4.3.7 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. In this study, the IIS approach was adapted to incorporate the RF and the ANN models.

The algorithm is divided in three steps (GALELLI; CASTELLETTI, 2013): (i) the IIS algorithm runs an input ranking algorithm to sort the variables with a nonlinear statistical measure of significance; (ii) the first p variables in the ranking are individually used as the input to a model building algorithm, so p single-input-single-output (SISO) models are constructed, and their performance is evaluated with a suitable metric; the best performing model is added to the final selection of input variables; (iii) the selected variables are used as an input to the model building algorithm multi-input single-output model (MISO) 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 two 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^2. Although the original IIS approach uses a model-free input ranking algorithm, here, the RF model was chosen, with the IncMSE as the significance measure, to be consistent with the first step of the methodology. The parameters for RF are the same from the first section of the methods. Although this strategy might slow down the algorithm, it still provides the desired ability of detecting nonlinear relationships and handling 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 of folds in the cross-validation was 2, 5 and 10 and ε varied between 0 and 0.1, with an incremental value of 10^{-2}.

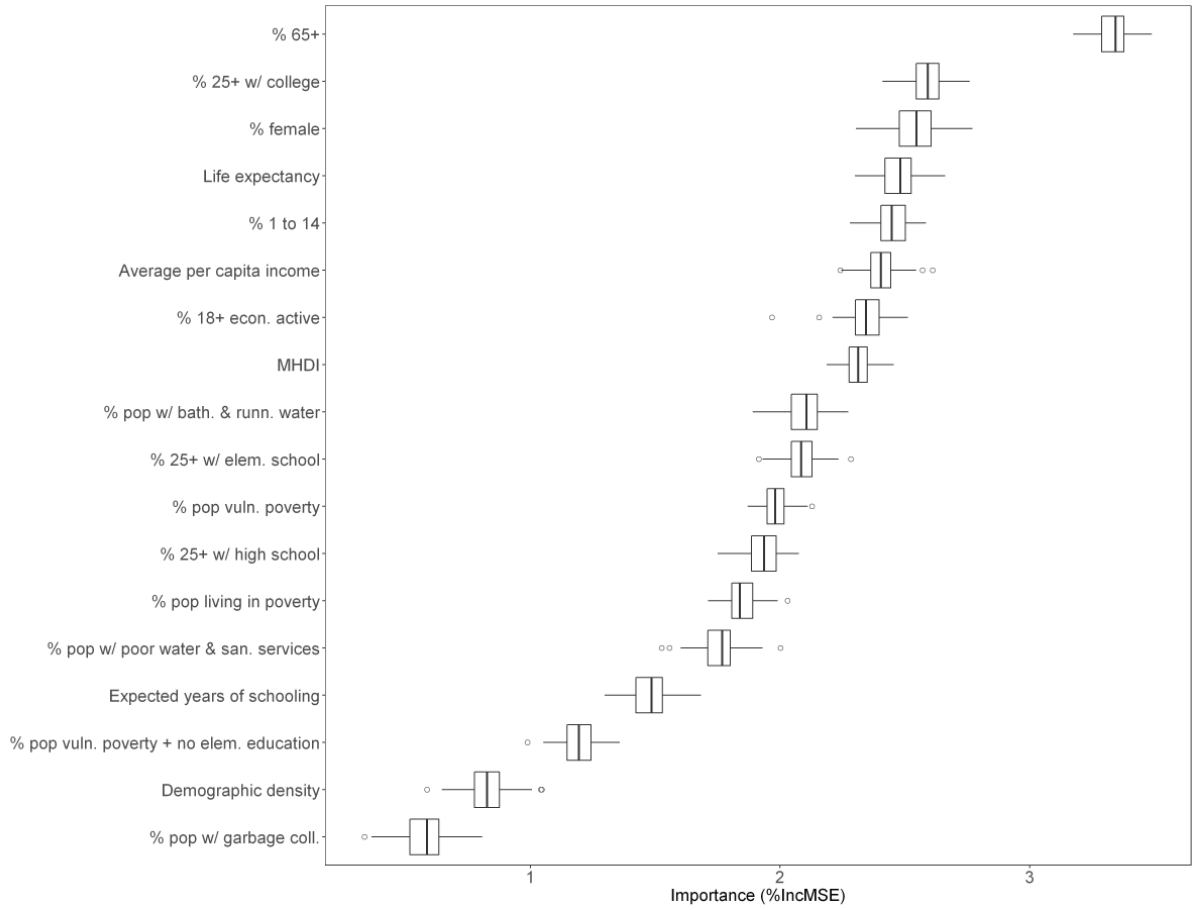
4.4 Results and discussion

4.4.1 Variable importance

The variables were ranked according to the median of the increase in MSE for 100 runs of the RF model (Figure 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 is irrelevant for model interpretation.

Figure 4 – Variable importance according to RF. The boxplots represent the variation in the average %IncMSE for 100 runs of the model. The variables are ranked according to the median value of the importance measure. See Table 5 for the description of the explanatory variables.



Source: The author.

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 were also 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. %65+ and %female, %1 to 14 and %25+ w/ college, % 25+ w/ college and Life expect.

The significance of household composition for water demand forecasting is corroborated by several studies (HOUSE-PETERS *et al.*, 2010; BENNETT *et al.*, 2013; MATOS *et al.*, 2014; HUSSIEN *et al.*, 2016; VILLARIN; RODRIGUEZ-GALIANO, 2019). Life expectancy and the presence of indoor bathrooms and running water might be useful to assess quality of life.

Besides, 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 (Figure 5). The average per capita income has a strong positive effect on the prediction. The influence of income in water use has been extensively explored in other studies (HOUSE-PETERS *et al.*, 2010; SHANDAS; PARANDVASH, 2010; LIU *et al.*, 2015; VILLARIN; 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; STAVINS, 2009). High-income households are less likely to be concerned about saving water than low- and medium-income households, who tend to maintain a lower consumption to avoid water shortage.

Percentage of children and elderly have opposite effects on water demand. The average prediction rises with increasing percentage of elderly (when above 4%) but falls with increasing percentage of children. An inverse relationship between households with children and water demand was also found in previous studies (SCHLEICH; 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; NOSVELLI, 2007), where water use tends to decrease with age. A positive relationship between elderly percent and the predictions could imply an increase in water demand in the next twenty years, since a demographic trend of population ageing is expected in Fortaleza (BARRETO; MENEZES, 2014)).

Some of the variables have a most significant effect on prediction after reaching a threshold, such as female percent, life expectancy, MHDI and percentage of adults which completed college and elementary school. The effect of the presence of bathrooms and running water in the households in average water demand is more significant between 88% and 93%.

The variables which were down in the RF ranking have little effect on the prediction. An increase in garbage collection coverage from 96% to 98%, for example, reduces average per capita water demand by only one unit. Some of these predictors have a complex relationship with the outcome and are difficult to interpret, such as exp. years of schooling, %pop vuln. poverty + no elem. education and dem. density.

4.4.2 Spatial analysis of water demand

After removing 45% of the least important variables from the ranking provided by RF, the ten remaining sociodemographic variables at the CB level were used to cluster water demand using SOM. The variables at the CT level (Human Development Index (MHDI) and per capita income) were also used to create clusters.

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

Table 6 – Characteristics of SOM clusters defined using the 10 most important explanatory variables at census block level. Except for area and population, the other variables are represented by the mean value for all census blocks in each cluster.

Cluster	A-CB (n = 6)	B-CB (n = 30)	C-CB (n = 55)	D-CB (n = 46)	E-CB (n = 45)
Total Area (km²)	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 (Lpd)	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$)	3622.77	1593.81	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

Source: The author.

The SOM map for CB data and its clusters are represented in the U-matrix (Figure 8). The heat maps in Figure 8 show the distribution of the explanatory variables across the SOM. They reveal a direct relationship between average per capita income, education level (%25+ w/ elem. school and %25+ w/ college), MHDI and economically active population percent. These have an inverse relationship with percentage of children. Female and elderly percent also present a direct connection.

The CB cluster's spatial distribution is represented in Figure 9 and their characteristics are listed in Table 6. Neighborhoods with high HDI and elevated per capita income were clustered together (A-CB and B-CB). These are also the areas with the highest water consumption rates. Further comments on the cluster divisions are provided in the supplemental material.

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

Table 7 – Characteristics of SOM clusters defined using the explanatory variables at census tract level. Except for area and population, the other variables are represented by the mean value for all census tracts in each cluster.

Cluster	A-CT (n = 24)	B-CT (n = 204)	C-CT (n = 128)	D-CT (n = 2596)
Total Area (km²)	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 (Lpd)	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

Source: The author.

The heat maps of the CT level SOM show a direct relationship between MHDI and average per capita income (Figure 10). The clusters are less representative than CB level's (Figure 11), probably because only two variables were used to create them. Areas with elevated average per capita income and MHDI were assigned to clusters A-CT and B-CT, which also present an elevated water consumption (Table 7). Census tracts with medium water consumption were clustered in C-CT. Cluster D-CT, which holds almost 90% of the population, incorporated census tracts with low per capita income and water use.

In order to verify that clusters were a good representation of water demand patterns, the water demand in each census tract and census block was compared to the average water demand of their corresponding clusters and the relative error was calculated. The mean relative error for each cluster is presented in Tables 8 and 9. CT level clustering (finer scale) resulted in better separated clusters than CB's (coarser scale) but was worst for water demand assessment (higher relative errors).

Although clustering could be used to improve prediction, the ANN performance

Table 8 – Mean of relative error (%) between water demand in census blocks and cluster average water demand.

Cluster	Census blocks				
	A-CB	B-CB	C-CB	D-CB	E-CB
Mean relative error (%)	18.45	20.08	21.05	20.01	17.31

Source: The author.

Table 9 – Mean of relative error (%) between water demand in census tracts and cluster average water demand.

Cluster	Census tracts			
	A-CT	B-CT	C-CT	D-CT
Mean relative error (%)	53.85	63.38	58.34	43.27

Source: The author.

would be reduced since some clusters have very few data points (A-CB, for example, contains 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 less variables provided better separated clusters, but the coarse scale is more convenient for urban planning and water demand estimation.

4.4.3 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 10) indicated that the best performing selected models are those with five SISO models and ten folds in the cross-validation. 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 is presented in Figure 12 The first two variables selected with IIS (av. per capita income and %1 to 14) were at the top of the RF ranking, while the third (%pop. living in poverty) had a rather low score. These three variables can fairly describe water demand in Fortaleza, with the av. per capita income functioning as a proxy for socioeconomic aspects of the households, %1 to 14 describing demographic aspects, and %pop. 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 are presented in Table 11. The results show that the CT model had a slightly better performance

Table 10 – Sensitivity analysis for the parameters of the Iterative Input Selection method. Each value represents the R^2 of the resulting model corresponding to the different parameters p, k and ε indicated.

ε	p = 1			p = 5			p = 10		
	k = 2	k = 5	k = 10	k = 2	k = 5	k = 10	k = 2	k = 5	k = 10
0	0.153	0.178	0.194	0.240	0.293	0.340	0.319	0.283	0.316
0.01	0.153	0.178	0.194	0.240	0.310	0.339	0.319	0.283	0.316
0.02	0.153	0.178	0.194	0.240	0.283	0.339	0.286	0.283	0.316
0.03	0.153	0.178	0.194	0.240	0.283	0.339	0.286	0.283	0.316
0.04	0.153	0.178	0.194	0.240	0.251	0.302	0.248	0.283	0.283
0.05	0.153	0.178	0.194	0.240	0.251	0.302	0.248	0.283	0.283
0.06	0.153	0.178	0.194	0.240	0.251	0.302	0.248	0.283	0.283
0.07	0.153	0.178	0.194	0.240	0.251	0.302	0.248	0.283	0.283
0.08	0.153	0.178	0.194	0.240	0.251	0.302	0.248	0.283	0.283
0.09	0.153	0.178	0.194	0.240	0.251	0.302	0.248	0.283	0.283
0.1	0.153	0.178	0.194	0.240	0.251	0.302	0.248	0.28	0.283

Source: The author.

than the CB model when comparing the R^2. One explanation is that the larger number of observations in the CT dataset benefits the training process of the MLP, which, as previously pointed out, requires large datasets. The ANN-CB model had only 182 observations, while the ANN-CT had 2952 and two independent variables.

Table 11 – Comparison of ANN-CB (three explanatory variables) and ANN-CT (two explanatory variables) model's performance.

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

Source: The author.

Water use patterns can differ depending on the aggregation level, since households with very different consumptions could end up in the same group. Bolorinos *et al.* (2020) also found that ML models perform better in a finer spatial scale. They showed that random forest 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.*, 2014). However, this study applied a linear model (linear mixed-effects and ordinary least squares regression), which has better performance when more spatial homogeneous data is used. For machine learning methods, the amount of data is determinant to the results, thus aggregating information might

reduce the learning power of the model. The influence of dataset size and the number of variables for ANN models was also pointed out by Lee and Derrible (2020), which showed that fewer explanatory variables are preferred when considering the same dataset size.

It is worth mentioning that in terms of R^2, both predictive models were only able to explain part of the residential water demand. Even at the CB scale, where 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. There are, however, 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. Also, 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.

4.5 Conclusion

In this study, three ML techniques were used to assess urban water demand in Fortaleza, Brazil: Random Forest, Self-Organizing Map and Artificial Neural Network. Two spatial levels were addressed: CB – coarse scale and CT – fine scale. The first had 18 sociodemographic explanatory variables, while the second had only two. 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 RF and ANN, was used to choose the best input variables for predicting water demand.

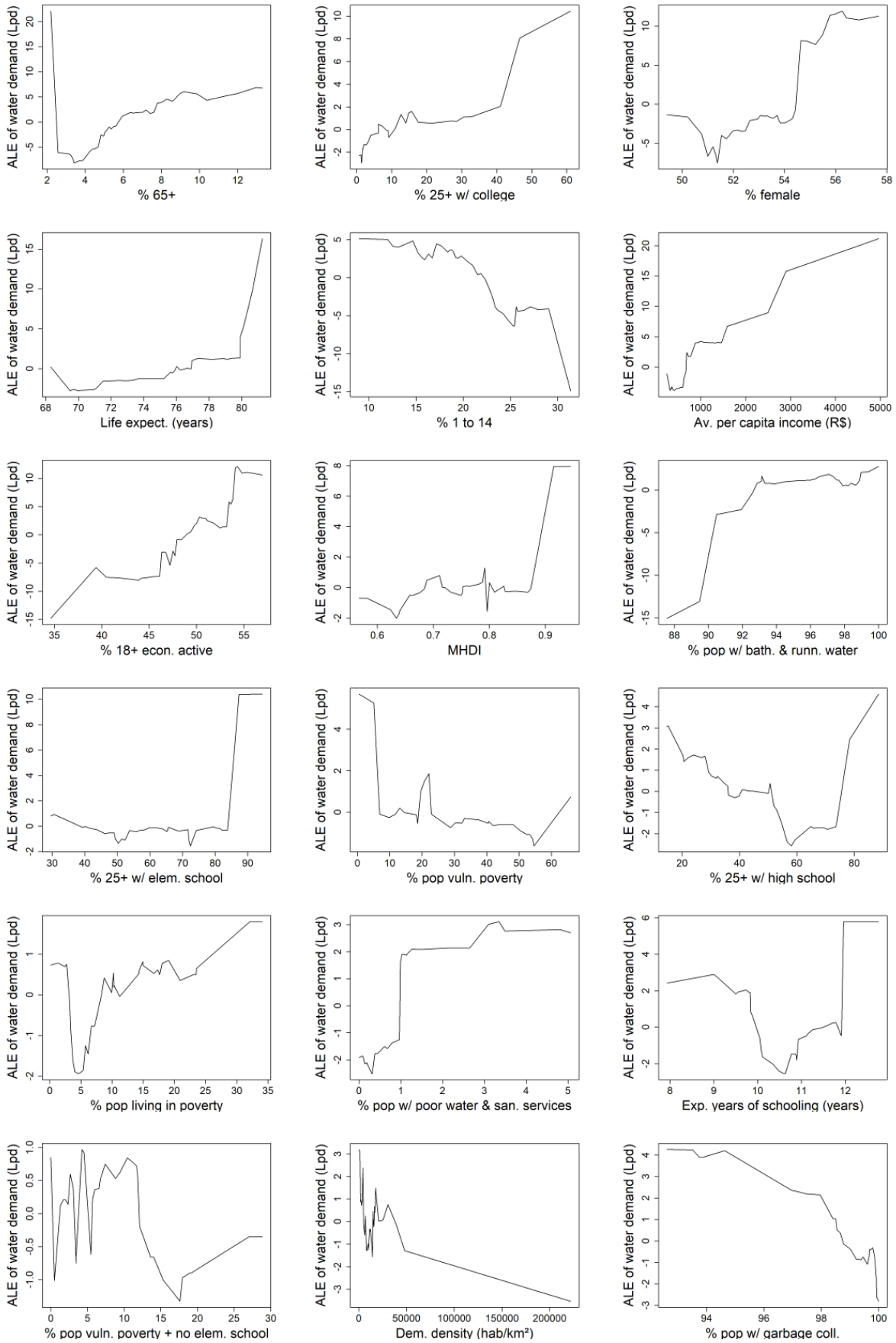
The features with the highest importance included those related to household composition (%65+, %female and %1 to 14), percent of college graduated inhabitants and life expectancy. The clustering analysis with self-organizing maps provided some interesting insights on the socioeconomic heterogeneity of Fortaleza. There is a distinct spatial gradient across the city regarding sociodemographic characteristics and water demand: central and eastern zones, with high water demand, have better education and health conditions, while southern and western regions, with reduced water demand, have low per capita income and MHDI. Nonetheless, heterogeneities in water demand are present inside 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 av. per

capita income, %1 to 14 and %pop. vulnerable to poverty provide a fair explanation of water demand in Fortaleza.

The aspects influencing water consumption are still not completely understood and ML methods are useful for identifying behavior patterns. Data availability has strong influence on the best approach for the modeling. If the dataset consists in high-dimensional data (in terms of 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.

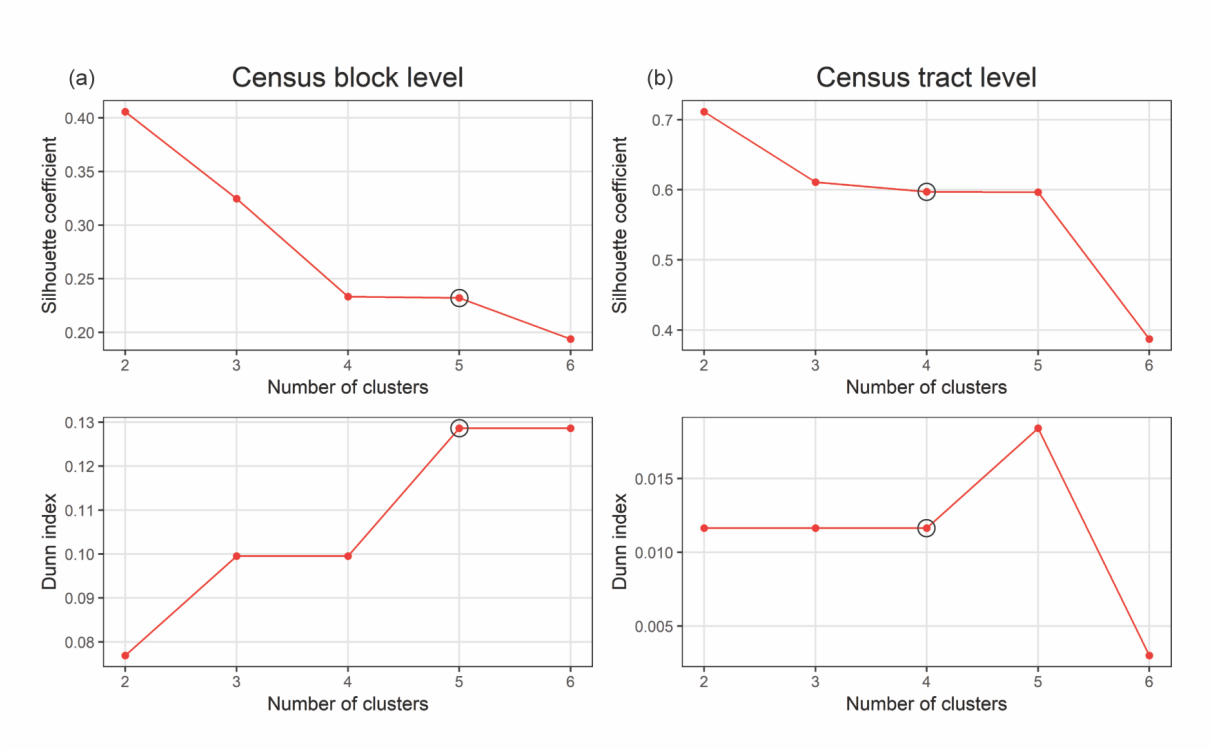
This study provided a better understanding on the influence of socioeconomic variables on the water demand of Fortaleza. The results are important not only for prediction, but also for designing targeted water conservation or pricing policies. Further studies could address temporal changes of water demand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

Figure 5 – Accumulated local effect plots for the RF model.



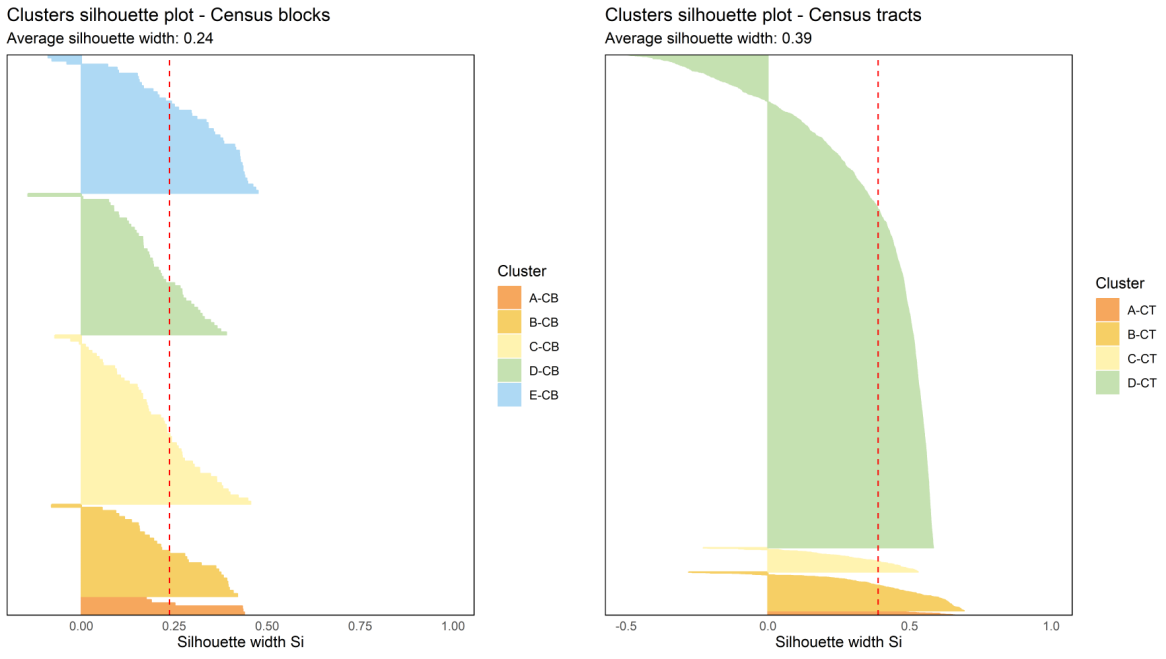
Source: The author.

Figure 6 – Dunn index and silhouette index for different number of clusters at the (a) census block and (b) census tract levels. The chosen number of clusters for each model are indicated with a black circle.



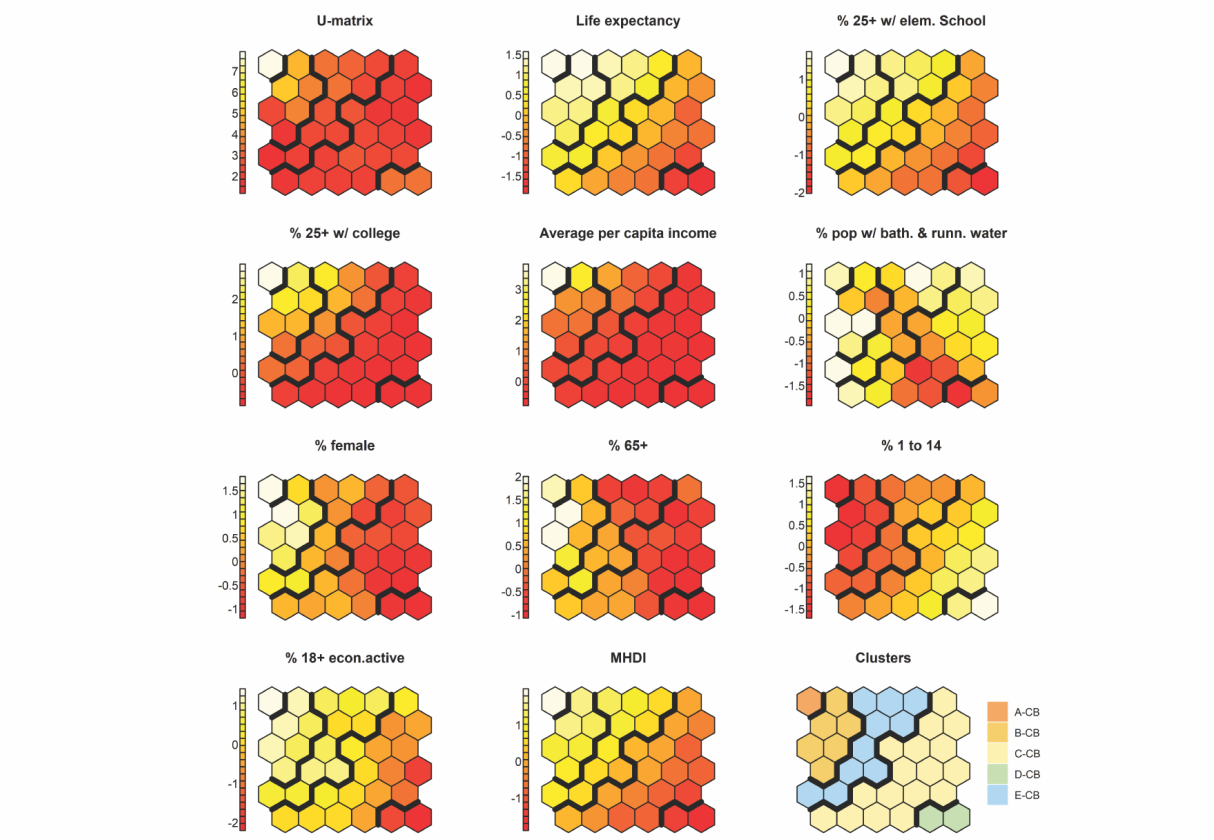
Source: The author.

Figure 7 – Clusters silhouette plot for census blocks (left) and census tracts (right) aggregation. For each census block or census tract, the figures show a straight horizontal line representing the silhouette coefficient. Each object is colored according to the correspondent cluster and the dashed red line represents the average silhouette width.



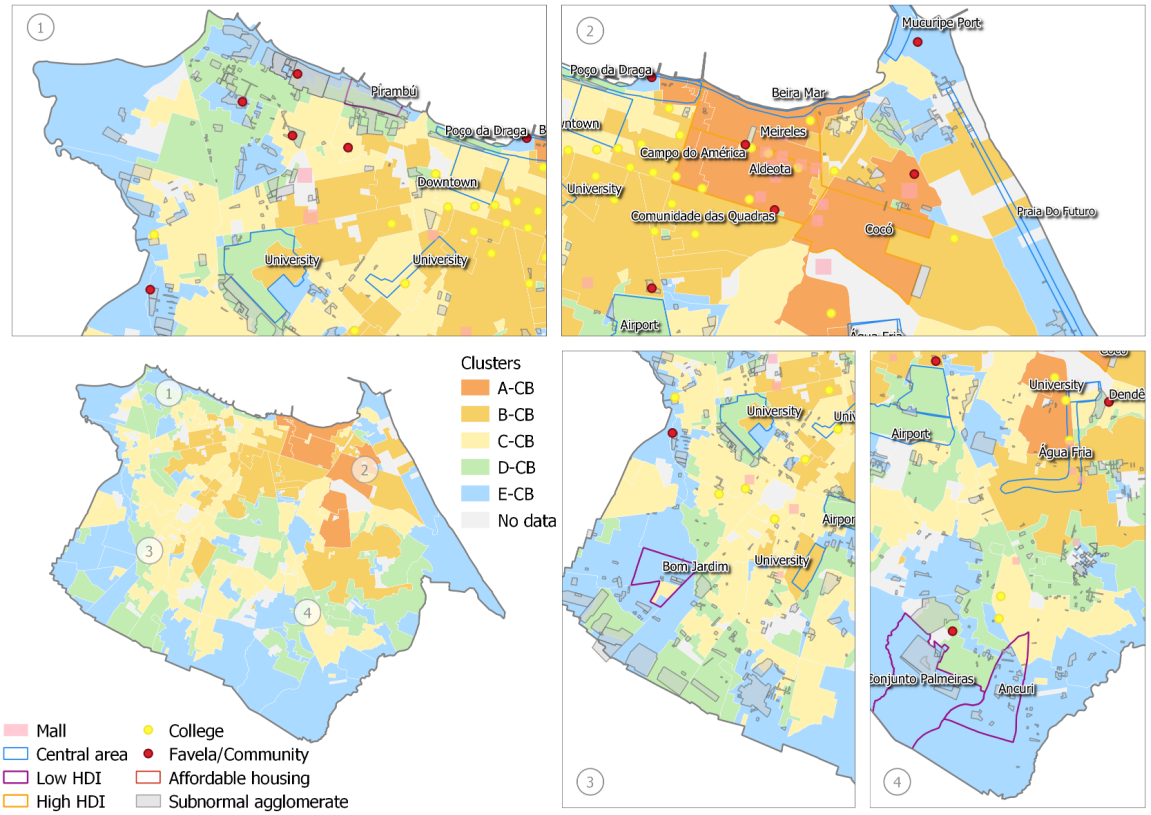
Source: The author.

Figure 8 – SOM heat maps for explanatory variables at census block level. The color gradient represents the Euclidean distance between each node and its neighbors, where light yellow means large distances and dark red small distances. See Table 5 for the description of the explanatory variables.



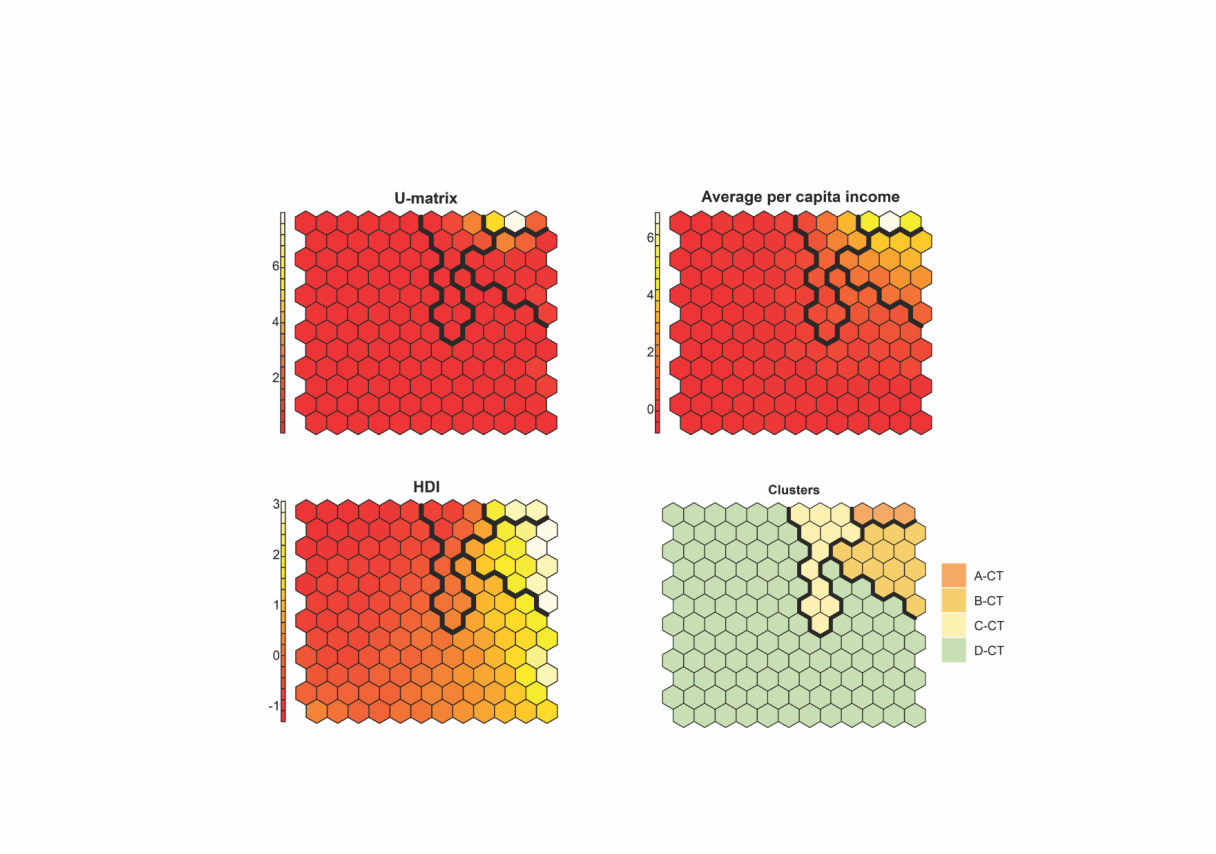
Source: The author.

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



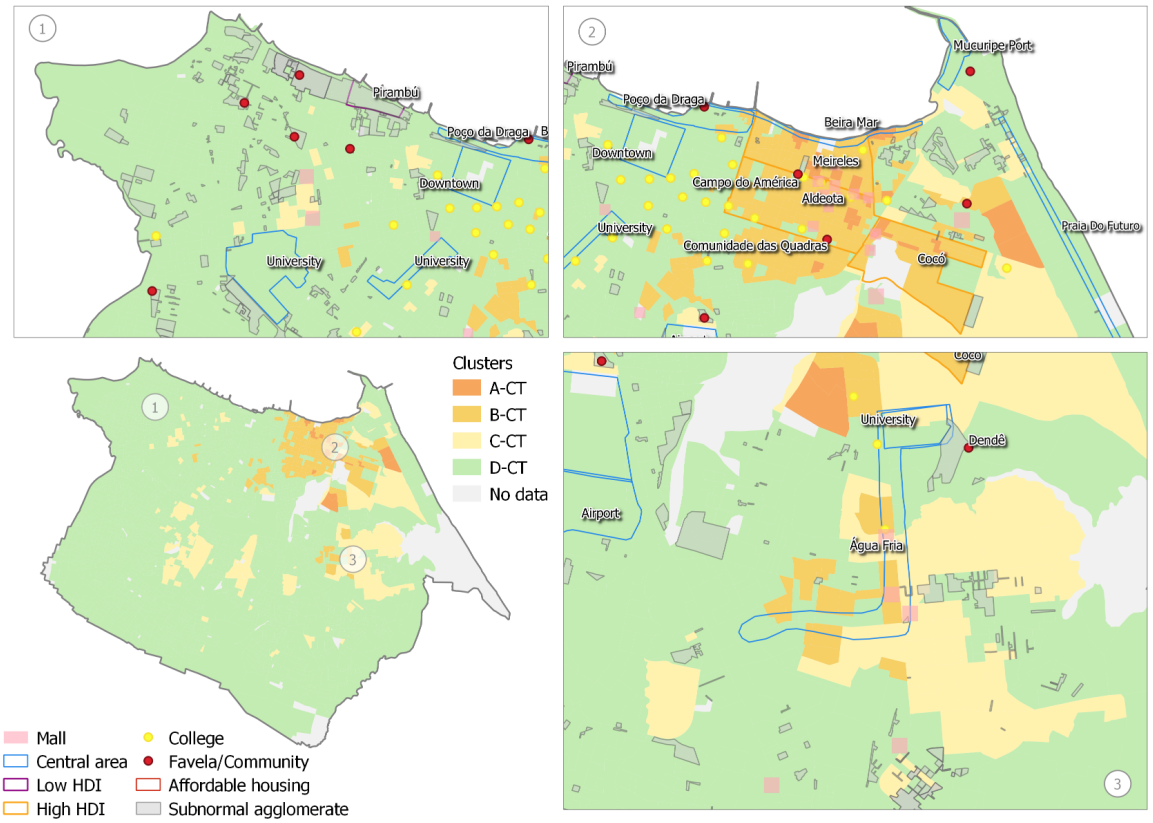
Source: The author.

Figure 10 – SOM heat maps for explanatory variables at CT level. See Table 5 for the description of the explanatory variables.



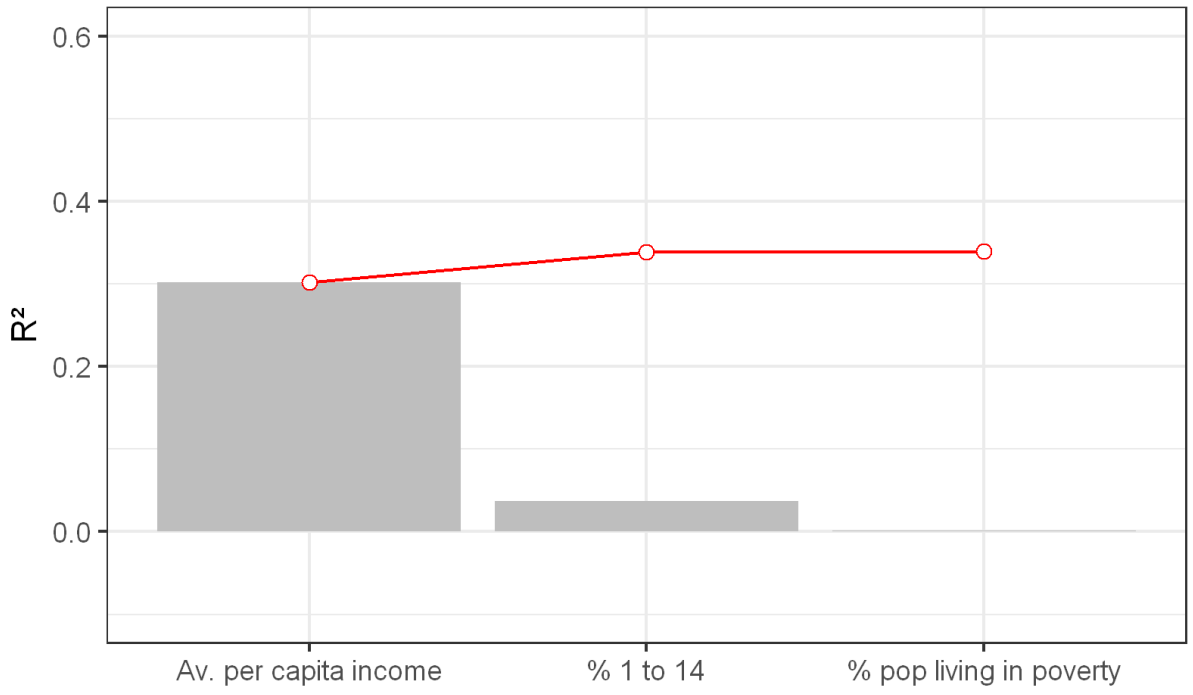
Source: The author.

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



Source: The author.

Figure 12 – Increase in performance (R^2) by adding the variables chosen in the IIS. The bars represent the increase in the R^2 obtained by adding each variable to the input dataset, while the red line represents the cumulated R^2.



Source: The author.

5 VARIATIONAL MODE DECOMPOSITION HYBRIDIZED WITH GRADIENT BOOST REGRESSION FOR SEASONAL FORECAST OF RESIDENTIAL WATER DEMAND

> A gente vive o repetido, o repetido, e, escorregável, num mim minuto, já está empurrando noutro galho. Acertasse eu com o que depois sabendo fiquei, para de lá de tantos assombros... Um está sempre no escuro, só no último derradeiro é que clareiam a sala. Digo: o real não está na saída nem na chegada: ele se dispõe para a gente é no meio da travessia. (ROSA, 2019)

5.1 Introduction

A primary concern of climate change and variability is how they will affect water demand and availability in the next decade (MILLY *et al.*, 2008; CISNEROS *et al.*, 2014). Spatial and temporal variability of precipitation and temperature might cause changes in the intensity and frequency of extreme events (ORLOWSKY; SENEVIRATNE, 2012). In urban systems, there is also the additional challenge of increasing urbanization and water use. Water resources planning should address accurate prediction of water demand, whether the objective is to expand the capacity of the supply system or to implement water conservation measures (OLMSTEAD, 2014).

Accurate forecasting of residential water demand is of special importance for the decision-making process, as researchers have shown it to be correlated with climate (MAIDMENT; MIAOU, 1986; HOUSE-PETERS; CHANG, 2011; ADAMOWSKI *et al.*, 2013; CHANG *et al.*, 2014). Specifically, it presents an inverse relationship with precipitation and a direct relationship with temperature (HOUSE-PETERS; CHANG, 2011; ADAMOWSKI *et al.*, 2013). Many other elements influence water demand patterns, such as demographic, social, and economic aspects of households (CHANG *et al.*, 2017; CHU; GRAFTON, 2019b; VILLARIN; RODRIGUEZ-GALIANO, 2019; LEE; DERRIBLE, 2020). These variables are associated with water demand trends and are usually predicted with scenario-based simulations.

Past research has indicated that water demand is strongly dependent on past use (DUERR *et al.*, 2018) and that it can be predicted only one month in advance. However, they also concluded that medium- and long-term forecasts could be improved by adding covariates. Short-term water demand forecasting, i.e. hourly to daily forecast, has been well explored. Lee

and Derrible (2020) evaluated twelve statistical models for residential water demand prediction, including eight machine learning techniques; gradient boost regression outperformed all the models. In their study, two scenarios of data availability were compared, and the one with a higher number of socioeconomic and climate exogenous variables provided better predictions.

Several studies have explored climate influence on residential water demand (ADAMOWSKI *et al.*, 2013; PARANDVASH; CHANG, 2016; ZUBAIDI *et al.*, 2020; RASIFAGHIHI *et al.*, 2020; FIORILLO *et al.*, 2021) Parandvash and Chang (2016) used a structural time series regression model to assess the effect of climate change on per capita water consumption and projected an increase of up to 10% in the water demand of Portland, in the United States, for the 2035-2064 period. Adamowski *et al.* (2013), Zubaidi *et al.* (2020) used decomposition techniques - wavelet transform and singular spectrum analysis, respectively - to detect interactions between climate and water demand. They found that decomposing time series into different components is a useful approach for filtering relevant information from exogenous variables. Haque *et al.* (2014), Rasifaghihi *et al.* (2020) provided long-term probabilistic forecasts of urban water demand, considering future climate projections. Some authors have investigated the joint influence of weather and socioeconomic aspects of households on water consumption (FIORILLO *et al.*, 2021).

To the best of our knowledge, the current models in the literature are not able to address the influence of climate on the medium-term forecast of water demand in dry regions. Our objectives are to (i) remove low-frequency variability and noisy signals from temperature and precipitation time series, (ii) extract the seasonal component of water demand, and (iii) design a model able to predict residential water demand up to 12 months in advance, considering the influence of precipitation and temperature variability. We do this by using an innovative approach that combines an intrinsic and adaptive decomposition method coupled with a regression machine learning model and use Fortaleza, Ceará – a region frequently affected by drought – as a case study. The VMD method used in this study was designed to concurrently estimate the components of a signal and properly deal with noise (DRAGOMIRETSKIY; ZOSSO, 2014). VMD was applied to extract the seasonal component of water demand, removing the signals unrelated to climate variability, and relevant signals from temperature and precipitation time series. Gradient boost regression was employed to capture the relationship between filtered signals of water demand and climate, which is long known to be nonlinear (MAIDMENT; MIAOU, 1986).

The study offers some important insights into tactical decisions on urban water

supply planning. The predictive model can be coupled with seasonal climate forecasts to assess future water demand and to guide the decision-making process.

5.2 Study area and data

The city of Fortaleza was used as a case study for the proposed model. Fortaleza is in the Northeast region of Brazil and is the fifth most populated city of the country, with over 2.6 million inhabitants. The region suffers from high climate variability and recurrent droughts, directly affecting Fortaleza's water supply. The most recent drought lasted seven years, starting from 2012 until 2018 (Pontes Filho et al. 2020). The rainy season occurs between February and May (Figure 13) and the maximum temperature ranges from 30 to 33 °C during the year (Figure 14).

Monthly residential water demand data from 2009 to 2017 was provided by the Water and Wastewater Company of Ceará. Data was provided at the household level, in cubic meters per month, and it was averaged over the number of consumers. Precipitation and maximum temperature time series were obtained from a conventional meteorological station maintained by the Brazilian National Meteorology Institute.

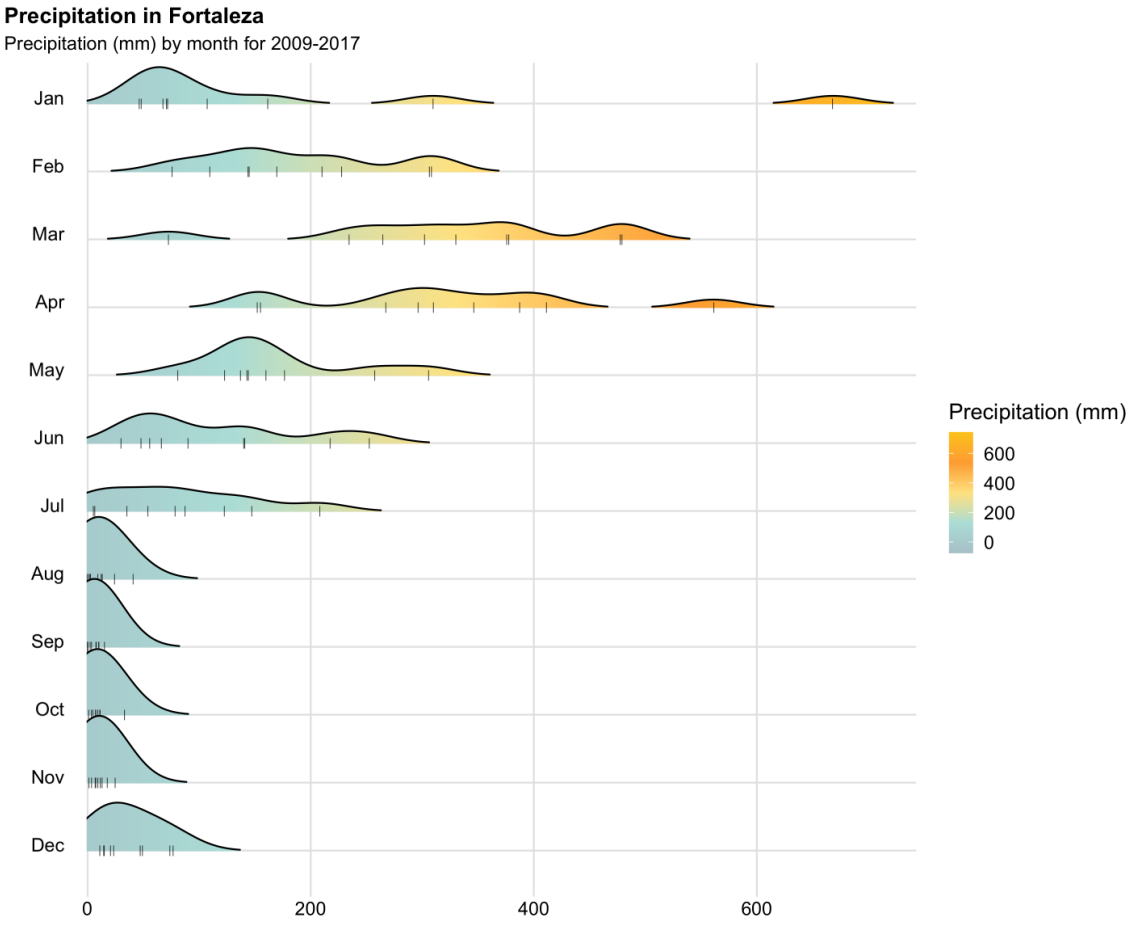
5.3 Methodoloy

5.3.1 Variational mode decomposition

Signal decomposition is a useful approach for filtering and capturing information from time series. The empirical mode decomposition (EMD) (HUANG *et al.*, 1998) is a famous time-frequency analysis used to process nonstationary and nonlinear series. Although this technique is simple and robust, there are a few limitations, such as the mode mixing problem, due to intermittent signals and noise, and the endpoint effect (GAO *et al.*, 2008). In addition, EMD lacks an appropriate mathematical theory basis. Some methods have been developed to solve these problems, such as the ensemble empirical mode decomposition (EEMD) (WU; HUANG, 2009), the complementary EEMD (YEH *et al.*, 2010), and the complete EEMD with adaptive noise (TORRES *et al.*, 2011). However, they were not able to address the mode mixing issue for all signals.

The VMD is a non-recursive decomposition method developed by Dragomiretskiy and Zosso (2014) to properly address the sensitivity to noise and sampling of EMD. The VMD

Figure 13 – Monthly average precipitation in Fortaleza for the period between 2009 and 2017. The rug plot represents original data points.



Source: The author.

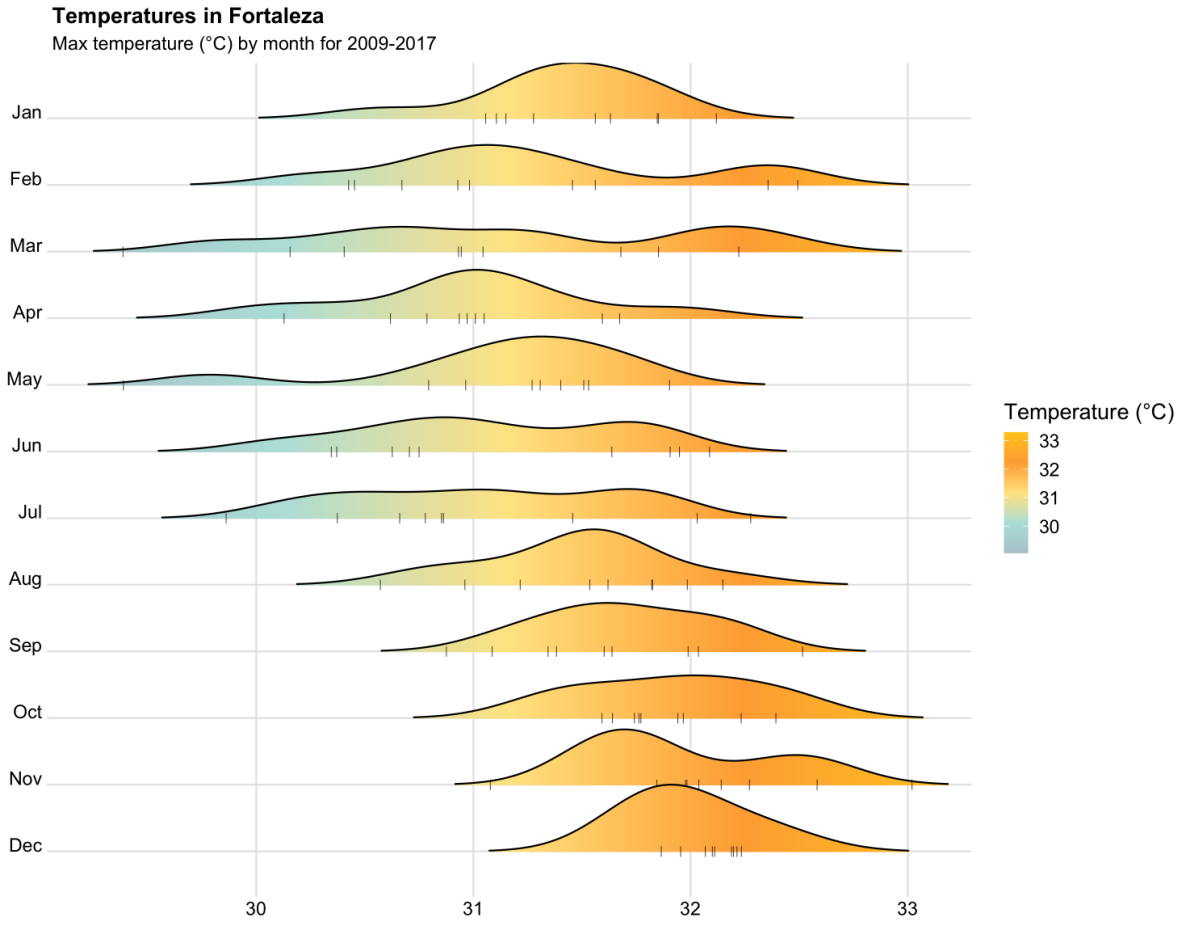
algorithm decomposes a signal into intrinsic mode functions (IMFs), which are amplitude-modulated frequency-modulated signals. Each mode is assumed to be compact around its center frequencies and they are concurrently estimated. The constrained variational problem solved by VMD to decompose a time series is given by the following equation:

$$min_{\{u_k\},\{\omega_k\}}\{\sum_k \|\partial_t[(\delta(t)+\frac{j}{\pi t}*u_k(t))]e^{-j\omega_k t}\|_2^2\}\ s.t. \sum_k u_k = f$$

where $\{u_k\}$ are the estimated modes, and $\{\omega_k\}$ their center frequencies, k is the number of IMFs, δ is the Dirac function, t is the time, $j^2 = -1$ and $*$ denotes convolution. For a complete description of the algorithm, see Dragomiretskiy and Zosso (2014).

VMD has three main parameters: the number k of IMFs, the quadratic penalty term α, and the convergence tolerance ε. To find the parameter k, we followed the approach suggested by Zuo *et al.* (2020), which is based on the observation of the center frequency of the last intrinsic mode function (IMF). After defining an initial value for k, we look at the amplitude spectrum; if

Figure 14 – Monthly maximum temperature in Fortaleza for the period between 2009 and 2017. The rug plot represents original data points.



Source: The author.

this decomposition mode presents the aliasing phenomenon, k is reduced by one and the analysis is repeated. A sensitivity analysis was performed to choose the best values for the quadratic penalty and the tolerance.

5.3.2 Gradient boosting regression

Gradient Boosting is a statistical model for function estimation based on a sequential ensemble of weak learners (FRIEDMAN, 2001). In this method, the weak learner – usually a decision tree – is first used to predict an output variable y with a set of explanatory variables x. Then, the weak learner (g_n) is used to predict the residuals of the initial model, and this procedure is repeated until the loss reaches a threshold or a maximum number of models is built (N). Predictions are multiplied by a learning rate or shrinkage parameter ν to slow down the

procedure and to increase the number of weak learners in the model:

$$f_n(x) = \nu * g_n(x)$$

The learning rate can vary between 0 and 1 but usually ranges from 0.1 to 0.3 (or less). The predicted value is added to the output of the previous model:

$$F_n(x) = F_{n-1}(x) + f_n(x)$$

Loss is minimized following a functional gradient descent algorithm. For regression tasks, the usual loss function is the mean squared error:

$$L(f) = \frac{1}{2}(y - F(x))^2$$

The gradient descent algorithm is used to optimize the parameters of the predictive model by finding the local minimum of the loss function:

$$f_n(x) = -\frac{\partial L(f)}{\partial F}$$

The main parameters of the gradient boosting model are: (i) the number of trees, which defines the number of iterations; (ii) the tree depth, which influences the complexity of the tree; (iii) the learning rate, and (iv) the minimum number of observations in a node to result in splitting. In this study, we set the learning rate to 0.1 and the number of observations per node to 10. We tested different combinations of the tree depth (1, 2, and 3) and the number of trees (50, 100, and 150). The model parameters were tuned using 5-fold cross-validation: the combination of parameters that provide the best performance across the cross-validation results is chosen.

5.3.3 Hybrid VMD-GBR model

To check the stationarity of the signals, the Augmented Dickey-Fuller (ADF) test was performed. This test assumes a unit root for the univariate time series, i.e., it tests the null hypothesis that $\alpha = 1$ in the following equation:

$$\Delta Y_t = c + \beta_t + y_{t-1} + \phi_1 \Delta Y_{t-1} + \phi_2 \Delta Y_{t-2} + ... + \phi_p \Delta Y_{t-p}$$

The inputs for the predictive model were selected using the mutual information (MI) between the signals of the weather variables and the filtered water demand and the partial

autocorrelation function (PACF) plots of each decomposed signal of water demand. The PACF approach is commonly used for streamflow forecasting (ALI *et al.*, 2020; FENG *et al.*, 2020). The confidence interval for the PACF corresponds to $[-1.96n, 1.96n]$, where n is the length of the training set; the significant lags are the ones that fall out of this interval.

The MI metric accounts for the interactions between two random variables without assuming linearity or continuity. Basically, the larger the value of MI, the closest the relationship between the variables and the amount of information that one contains about the other. MI is based on the concept of Shannon entropy, which measures the uncertainty of a variable. The MI between two variables X and Y is expressed as:

$$I(Y;X) = sum_{x \in X} sum_{y \in Y}(x,y) loglog(\frac{p(x,y}{p(x)p(y)}$$

The methodology of the VMD-GBR model can be summarized as follows:

Step 1: Decompose the water demand, precipitation, and maximum temperature time series into additive intrinsic mode functions using VMD. The parameter k is defined by observing the power spectrum of the last IMFs of each decomposed signal, which should not present a center frequency alias (ZUO *et al.*, 2020). The quadratic penalty term and the convergence tolerance are chosen with sensitivity analysis on model performance.

Step 2: Estimate the deterministic component of the signals of water demand using the ADF test and reconstruct the time series using only the remaining signals.

Step 3: Detect the most relevant IMFs of the weather variables by calculating the mutual information between each of them and the reconstructed signal of water demand. These will be inputs for the predictive model.

Step 4: In addition to the IMFs selected in the previous step, choose the lagged inputs for the predictive model by observing the partial autocorrelation function of the water demand IMFs. The IMF corresponding to the trend component is not included in this analysis.

Step 5: Normalize all data using the min-max normalization:

$$x_{norm} = \frac{x - min(x)}{max(x) - min(x)}$$

Step 6: Split the dataset into training and testing (here, we used 80% for model training and 20% for testing). The input variables are the lagged IMFs of water demand and the most relevant IMFs of weather variables. In this study, different combinations of the model parameters were tested, namely, the number of trees, the tree depth, shrinkage, and the number of

observations in the terminal nodes. The parameters are tuned using 5-fold cross-validation in the training dataset and the model performance is evaluated using the testing dataset.

5.3.4 Performance assessment

Model performance was evaluated with three measures: R^2, MSE, and RMSE.

$$R_2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

$$MSE = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|^2}{n}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|^2}{n}}$$

where y_i is the observed water demand at month i, y_i is the predicted water demand at month j, and n is the number of months in the prediction horizon.

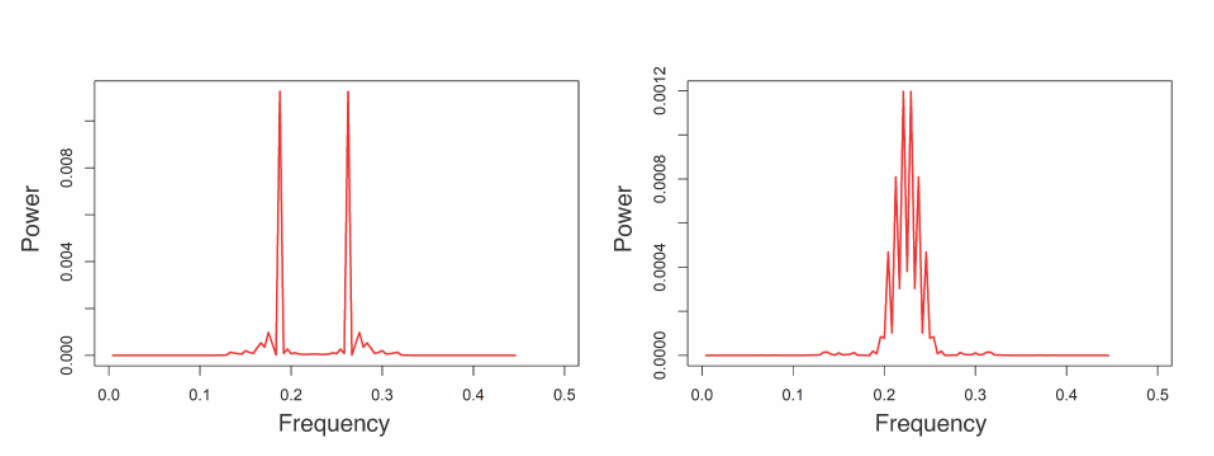
5.4 Results and discussion

The residential water demand time series was decomposed into four signals to avoid the aliasing effect observed in the last IMF when k was set to five (Figure 15 and Figure 16). Following the same approach, the precipitation and maximum temperature time series were decomposed into three IMFs each (Figure 17 and Figure 18).

The MI metric indicated that the second IMF of both maximum temperature and precipitation were the ones to contain the most information on the water demand series (Table 12). The autocorrelation functions of these signals present a seasonal pattern where the peaks and the troughs are six months apart, while the third IMF does not seem to have a seasonal pattern. This might indicate that the last IMF of each series contains noise and thus could not directly influence demand patterns, while the second corresponds to a periodic signal.

The second IMF of water demand decomposition corresponds to the trend component. The decreasing trend in residential water demand after 2015 could be associated with conservation attitudes. After the 2012-2018 drought, the local water company implemented a contingency tariff to encourage a reduction of at least 20% in consumption. Socioeconomic factors, such as

Figure 15 – Power spectrum of IMFs 4 (left) and 5 (right) of water demand time series. The aliasing effect can be observed in the IMF5, where the center frequency overlap.



Source: The author.

income, water price, and household composition could also be associated with changes in water demand trends, as pointed out in previous studies (PARANDVASH; CHANG, 2016; ZUBAIDI *et al.*, 2020). Demand-side measures and even mass media coverage of extreme events can also affect the behavior of this particular signal of water demand (BOLORINOS *et al.*, 2020). Modeling this component was beyond the scope of this study.

Table 12 – Mutual information between each decomposed signal and filtered water demand time series.

Max Temperature			Precipitation		
IMF_1	IMF_2	IMF_3	IMF_1	IMF_2	IMF_3
0.07	0.22	0.06	0.07	0.11	0.05

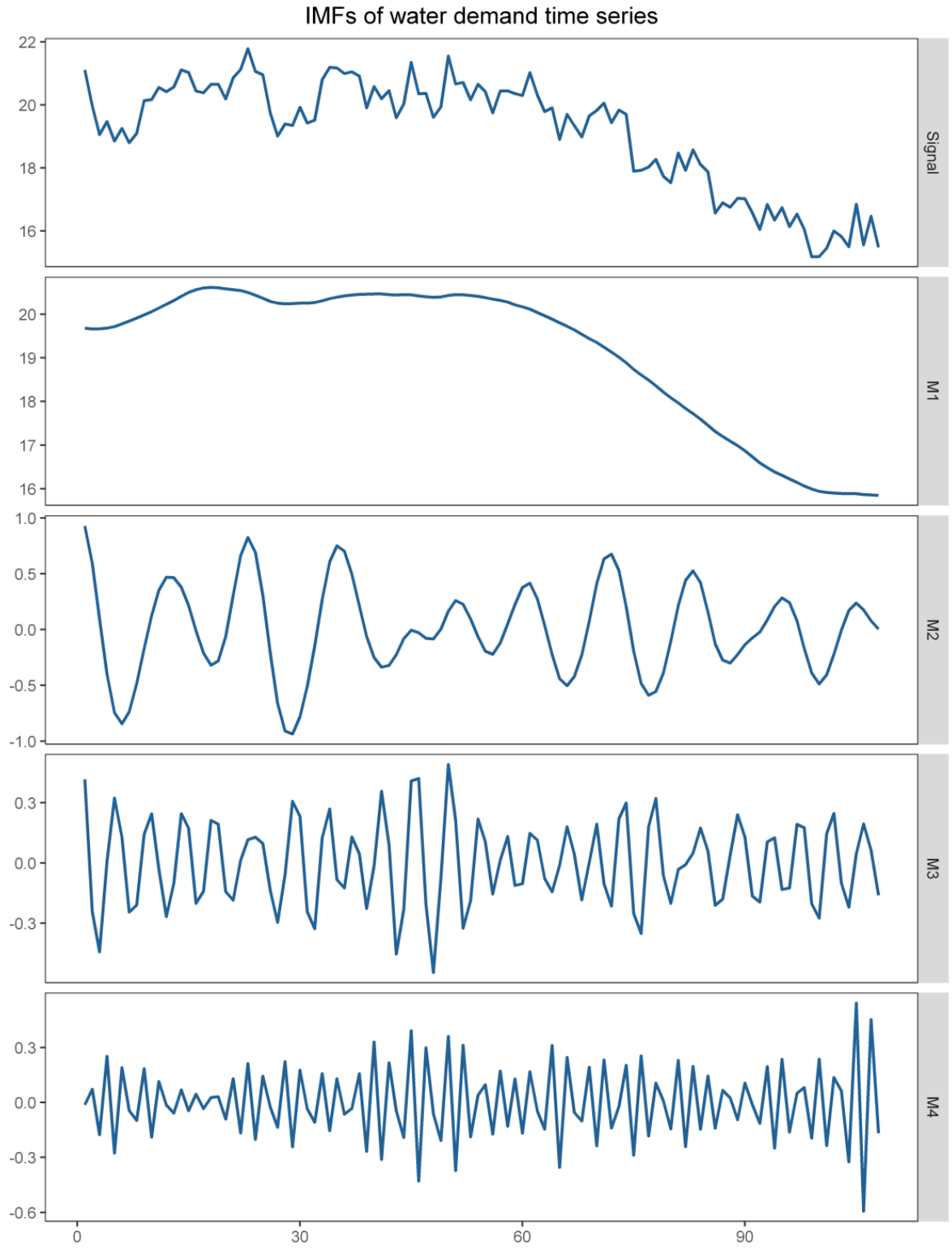
Source: The author.

The additional relevant inputs were defined based on the PACF of the decomposed signals of water demand (Figure 19). Previous water demand has a great influence on future consumption and climate variables alone would not be able to provide accurate predictions. The final dataset had 12 input variables.

A sensitivity analysis on the performance of the VMD-GBR model for 1-month ahead predictions indicated the most suitable values for the quadratic penalty term and the convergence tolerance, set to 10 and 10-5, respectively. Table 13 indicates the R^2 values for different combinations of both parameters. After defining these parameters, the model was tested for predictions with leading times varying between one and twelve months.

Figure 20 presents the scatter plots of the testing set for each leading time. As

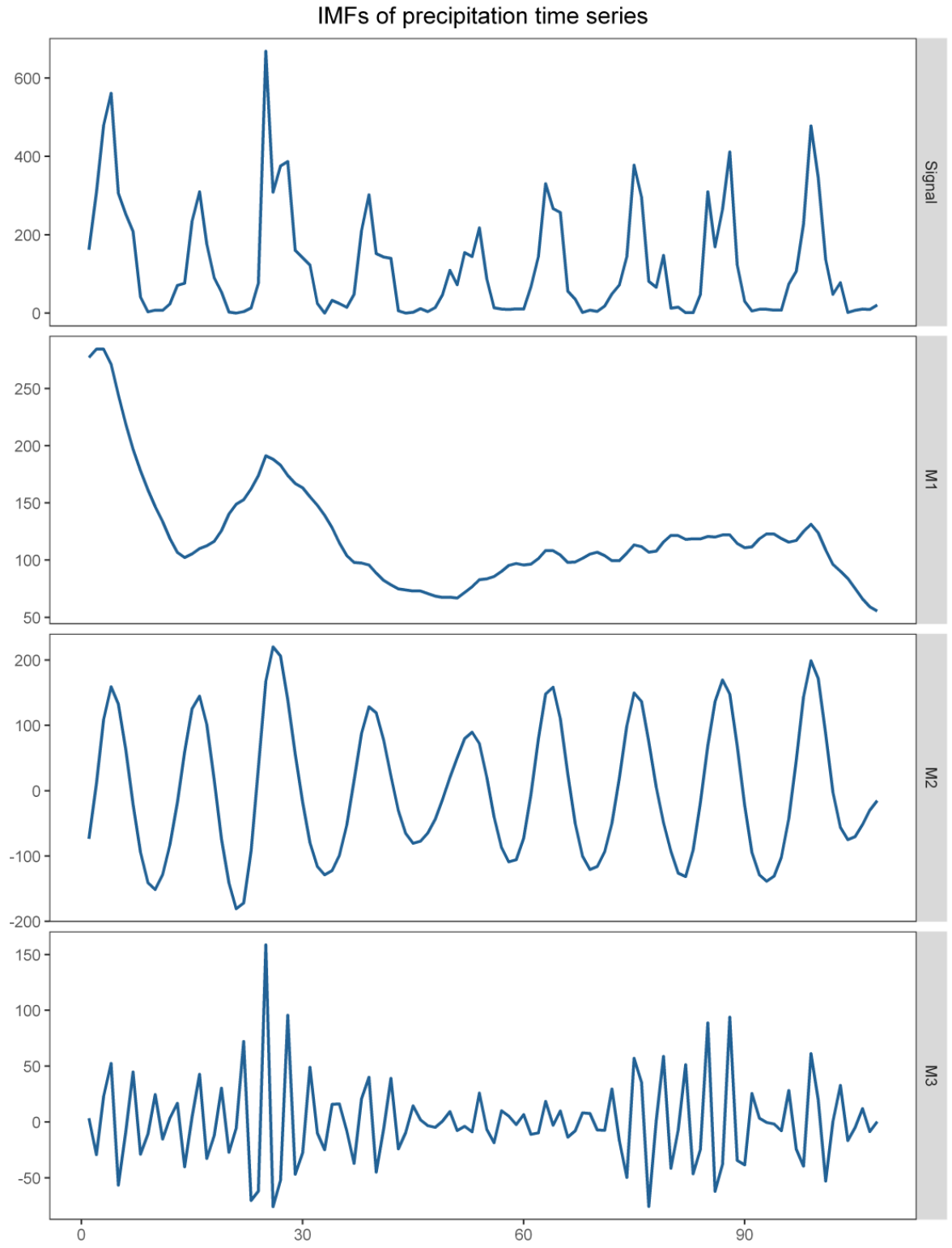
Figure 16 – Original and decomposed signals of water demand time series.



Source: The author.

it would be expected, the performance is worse as the leading time increases, but the model presents accurate predictions for 1, 2, 3, and 4-months ahead of water demand. Table 14 shows the R^2, RMSE, and MSE for each leading time. The VMD-GBR model successfully addresses

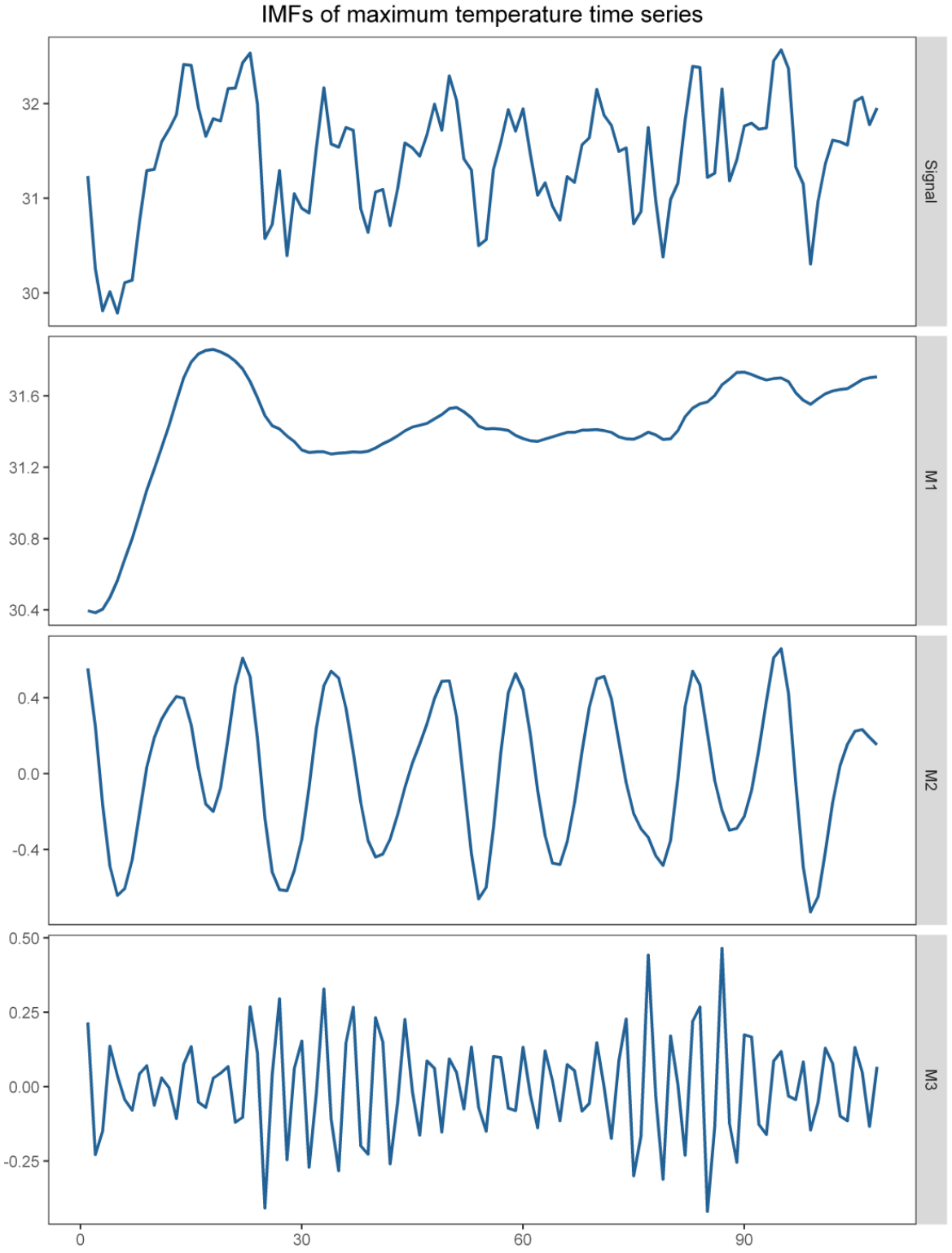
Figure 17 – Original and decomposed signals of mean precipitation time series.



Source: The author.

climate variability in water demand prediction and reassures previous findings that residential consumption is driven by precipitation and temperature patterns (ADAMOWSKI *et al.*, 2013; PARANDVASH; CHANG, 2016; ZUBAIDI *et al.*, 2020).

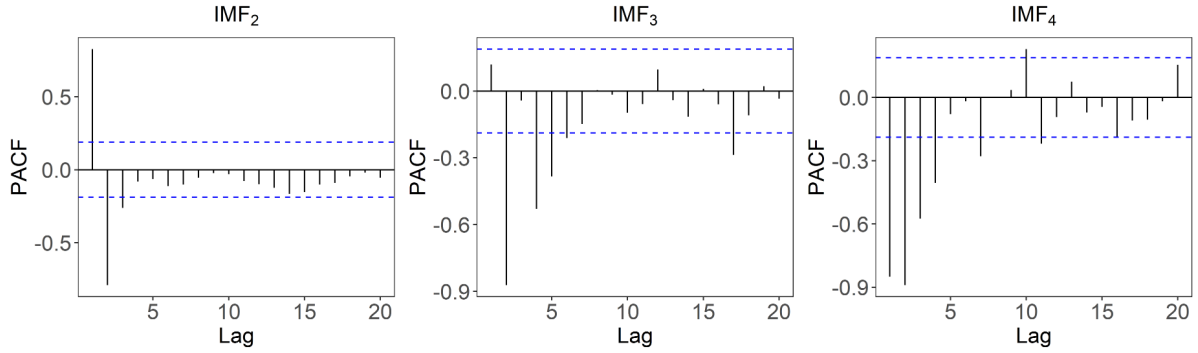
Figure 18 – Original and decomposed signals of maximum temperature time series.



Source: The author.

The importance measure of the input variables provides insight into the influence of climate variables in the prediction (Figure 21). Although there is a large variance in the mean average MSE of the IMFs of temperature (tmpIMF2) and precipitation (precIMF2), they are

Figure 19 – Partial autocorrelation plots of water demand IMFs.



Source: The author.

Table 13 – R^2 for different combinations of VMD parameters.

α	ε						
	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-12}	10^{-15}	0
10	0.719	0.714	0.705	0.705	0.705	0.705	0.705
20	0.680	0.697	0.697	0.697	0.697	0.697	0.697
50	0.711	0.700	0.700	0.700	0.700	0.711	0.700
100	0.675	0.675	0.675	0.675	0.675	0.675	0.675
200	0.710	0.710	0.710	0.710	0.710	0.710	0.717
500	0.323	0.323	0.323	0.323	0.323	0.323	0.323
600	0.307	0.307	0.307	0.307	0.307	0.307	0.307
700	0.276	0.276	0.276	0.276	0.276	0.276	0.276
800	0.282	0.283	0.283	0.283	0.283	0.283	0.283
900	0.273	0.273	0.273	0.273	0.273	0.273	0.273
1000	0.272	0.266	0.266	0.266	0.266	0.266	0.266
2000	0.192	0.185	0.185	0.185	0.185	0.185	0.185

Source: The author.

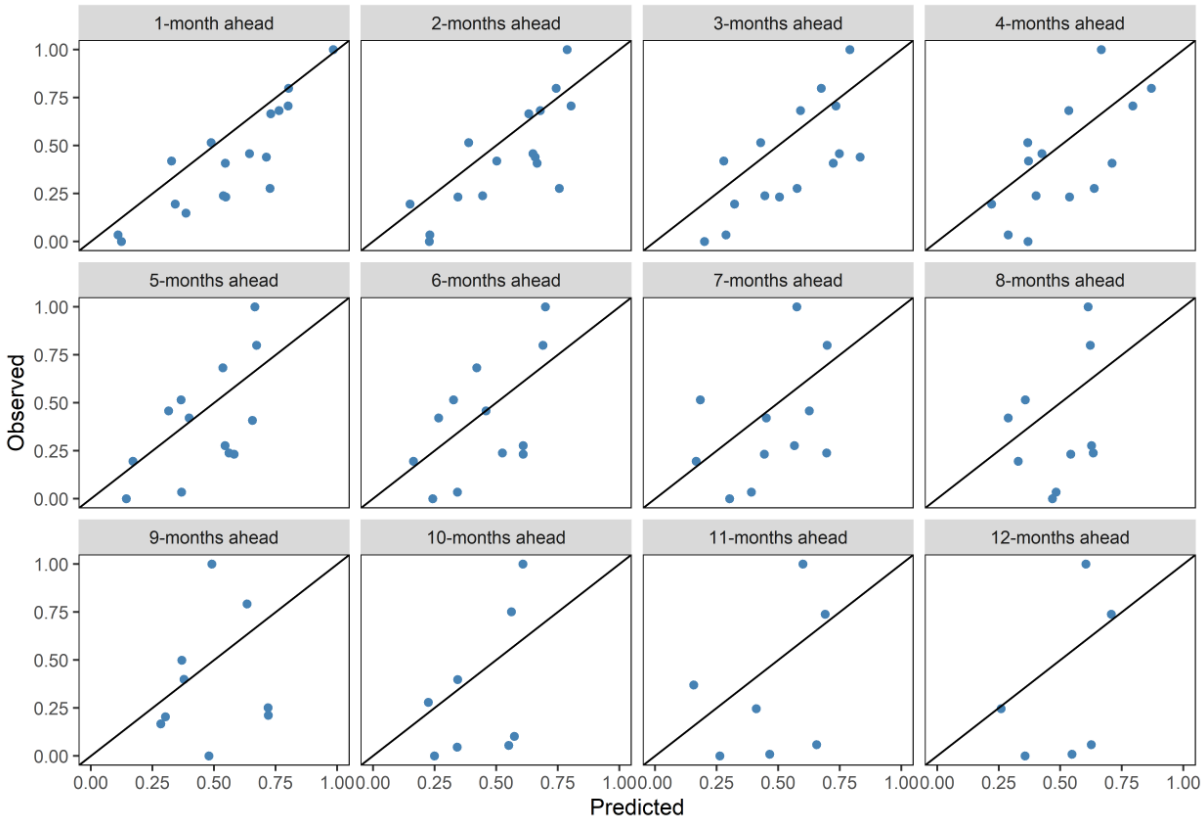
Table 14 – Performance metrics for the VMD-GBR model predictions during the testing period for different leading times.

Lead time (months)	R^2	RMSE	MAE
1	0.719	0.197	0.158
2	0.549	0.222	0.188
3	0.463	0.226	0.199
4	0.519	0.213	0.173
5	0.388	0.230	0.192
6	0.295	0.258	0.226
7	0.354	0.258	0.230
8	0.110	0.312	0.278
9	0.233	0.277	0.233
10	0.290	0.319	0.271
11	0.337	0.324	0.271
12	0.324	0.375	0.313

Source: The author.

amongst the top-ranked variables. This result confirms the hypothesis that residential water

Figure 20 – Scatter plots of the normalized fitted values of the VMD-GBR model and normalized observed data for the testing period for each leading time.



Source: The author.

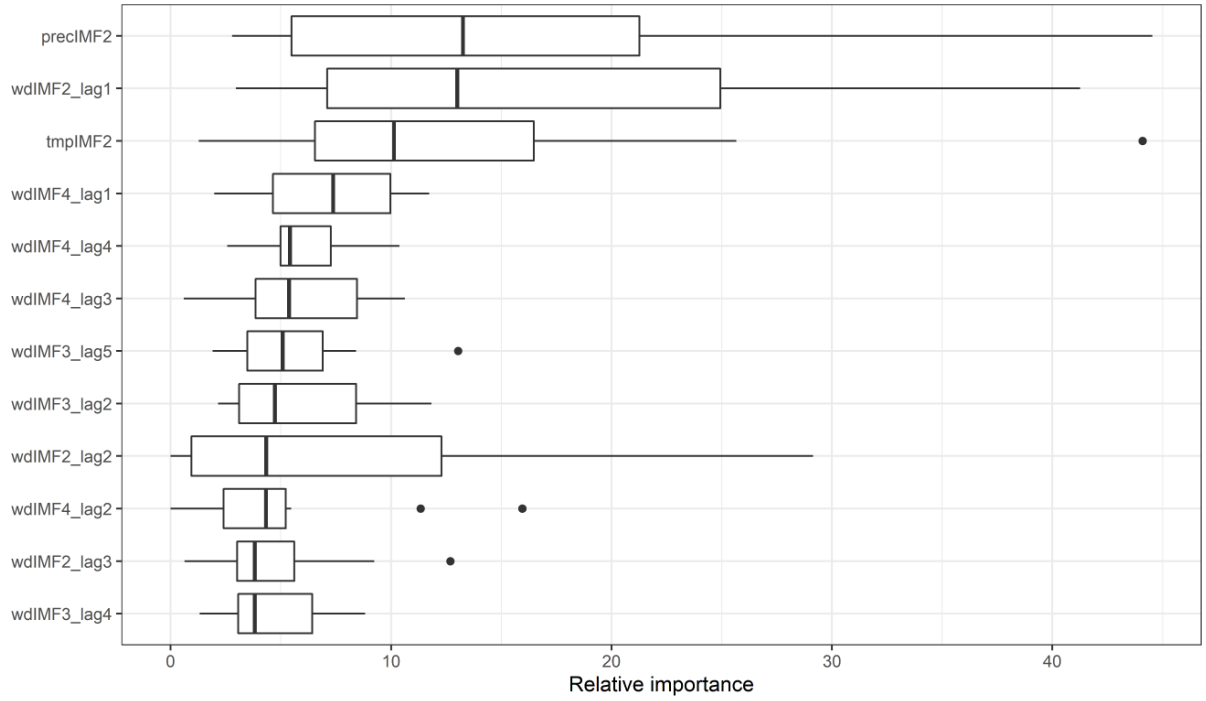
demand is driven by climate patterns.

Different from the application area of other researches mentioned here (PARANDVASH; CHANG, 2016; RASIFAGHIHI *et al.*, 2020; FIORILLO *et al.*, 2021), Ceará has a significant interannual variability of both precipitation and temperature, mainly due to the El Niño South Oscillation and the Interhemispheric Tropical Atlantic Gradient (HASTENRATH; HELLER, 1977). The region also presents intraseasonal variations related to the Madden-Julian Oscillation (JUNIOR *et al.*, 2018b). Although widely studied, these phenomena have complex interactions with precipitation that are still not completely understood by the scientific community. Hence, forecasting models that can properly detect seasonal variability of climate variables and their relationship with water demand can be of great value for operational management decisions and the adjustment of demand-side strategies.

5.5 Conclusion

This study set out to design a predictive model of monthly residential water demand including climate variability. To do that, we applied a decomposition technique to remove

Figure 21 – Boxplot of the increase in MSE obtained when each of the input variables was removed from the dataset, ranked according to the median value of its relative importance.



Source: The author.

the water demand component associated with socioeconomic and policy characteristics and a machine learning technique to create an autoregressive model. The methodology is applied in Fortaleza, Brazil, a region with an elevated interannual and intraseasonal climate variability.

The results show that applying VMD to filter the water demand signal is an effective approach for removing components that are not directly associated with climate variability. Although the trend component could be associated with a response to drought, that is somehow dependent on climate, the effective implementation of water conservation policies and the change of habits in the households are more related to socioeconomic factors. The VMD-GBR model is suitable for regions affected by extreme events or complex climate variability.

Maximum temperature and precipitation were significant predictors of water demand and including their seasonal components as exogenous variables of the model improved accuracy. The model is appropriated for at least 4 months-ahead predictions, with an average RMSE of 0.214. The methods used in this study may be applied to medium-term planning of water supply systems and to guide operational and tactical decisions of water companies. The VMD-GBR approach can yet be coupled to seasonal climate forecast models and scenario-based predictions of the trend component of water demand. The findings are also useful to assess climate change impacts on future water demand, which could provide insight into policy design.

6 A DATA-DRIVEN MODEL TO EVALUATE THE MEDIUM-TERM EFFECT OF CONTINGENT PRICING POLICIES ON RESIDENTIAL WATER DEMAND

> "Mas levei minha sina. Mundo, o em que se estava, não era para gente: era um espaço para os de meia razão." (ROSA, 2019)

6.1 Introduction

The growing water demand associated with urbanization processes has increased water stress and the risk of shortage in several regions of the world (MCDONALD *et al.*, 2014). For some of them, the elevated temporal and spatial variability in water availability offer an additional challenge to water supply management (ORLOWSKY; SENEVIRATNE, 2012; PAL *et al.*, 2013; CAMPOS *et al.*, 2014).

In this context, water companies and policymakers have been implementing demand control measures, since increasing water supply capacity is not always possible or effective (ROMANO *et al.*, 2014; WHITTINGTON; NAUGES, 2020). A widely used approach is the adoption of increasing block rates (IBR), which is expected to encourage rational water consumption (RIETVELD *et al.*, 2000; ZHANG *et al.*, 2017). This kind of policy is typical of regions affected by droughts and developing countries and has complex impacts on consumer behavior (RINAUDO *et al.*, 2012). Pricing strategies might also include tariffs that vary seasonally with temperature and/or precipitation (PESIC *et al.*, 2012; MOLINOS-SENANTE, 2014) or adjusted with the level of water storage (CHU; GRAFTON, 2019a), and household size (ARBUÉS; BARBERáN, 2012).

A less common strategy to reduce water use under drought conditions is the implementation of penalty fees for those households with an elevated consumption (GARCÍA-RUBIO *et al.*, 2015; BRAGA; KELMAN, 2020). In Brazil, water utility companies have used this approach to deal with water crisis (BRAGA; KELMAN, 2020). In Fortaleza, located in northeast Brazil, water pricing follows an IBR structure, and a contingent tariff, i.e., a penalty fee, was adopted three years after the beginning of a severe drought that reduced reservoir storage by about 63% (PONTES FILHO *et al.*, 2020). This tariff was influenced by the consumption quantity that exceeded a predefined threshold.

Previous studies have reported that water scarcity impacts price elasticity, but the consequences are adverse. While early research indicated that price elasticity is more significantly

affected by pricing structure and season (ESPEY *et al.*, 1997), recent studies show that consumers response to price change is related to different exogenous factors, such as climate (MONTEIRO; ROSETA-PALMA, 2011), income (MA *et al.*, 2014) and environmental attitude (GARRONE *et al.*, 2019). Dalhuisen *et al.* (2003) pointed out that income elasticities are relatively inelastic under IBR pricing, and that water scarcity does not seem to affect elasticity. Molinos-Senante and Donoso (2016) proposed a tariff scheme that accounts for the scarcity value of water and that can promote equity, based in a IBR structure and cross-subsidy. However, the measure might be difficult to implement due to lack of adequate water metering. Another strategy aiming equity and sustainability was presented by Ward and Pulido-Velázquez (2008), that presented a two-tiered pricing setup. Debate continues about the effectiveness of price control policies for demand control, especially on IBR schemes (MANSUR; OLMSTEAD, 2012; ZHANG *et al.*, 2017; MATIKINCA *et al.*, 2020).

The research to date has extensively explored the price influence on water consumption (ARBUéS *et al.*, 2004; OLMSTEAD *et al.*, 2007; WARD; PULIDO-VELáZQUEZ, 2008) – together with other socioeconomic and/or climatic variables – but only a few studies are able to address it over long time horizons (GRAFTON *et al.*, 2014). Most studies on water price use survey (which can be expensive and time consuming), aggregate, or household level data to assess the empirical implications of economic variables on water demand (RUIJS, 2009). Although these analyses have improved the understanding of the scientific community and decision-makers, they do not allow continuous learning as new data become available.

Water companies have a huge amount of smart meter data available that could be useful to extract information on use patterns and consumer behavior (COMINOLA *et al.*, 2019). In this research, we present a method that benefits from this data to support managers on how to adjust the pricing policy for a planning horizon of up to one year. The model can be coupled with reservoir/supply systems operation or water distribution models to provide further insights on supply-demand balance strategies.

This study proposes a data-driven predictive model to assess the medium-term effect of price-based water conservation policies at the household level. In addition, we calculate the elasticity of water demand reduction to price and we assess how much water price and public interest in the drought can affect consumption habits. The methodology can be used by water companies to assess price-related strategies of water conservation and does not require additional variables that could be difficult to obtain in a refined scale. An advantage of this model is

that prediction can be performed at a disaggregated level, making it possible to design policies tailored to socioeconomic or even structural characteristics of the households. Although this study considers a block tariff structure, the framework can be adapted to any other price strategy, if it is applied at the household level.

6.2 Methodology

6.2.1 Study area

The city of Fortaleza, capital of Ceará, located in the Northeast region of Brazil, is the fifth most populated city of the country, with over 2.6 million inhabitants distributed across $314.9 km^2$. The population is expected to grow to 3.1 million people in 2040 (IPLANFOR, 2015). The city is part of the Metropolitan Region of Fortaleza, which comprises 19 municipalities of Ceará.

Fortaleza is supplied by the JMS, which consists of eight reservoirs which sum up to a storage capacity of $11,112 hm^3$. JMS transfers water from the Jaguaribe basin and supplies 36 municipalities. Urban and industrial demand of Fortaleza is$6.77 m^3/s$, corresponding to 56.5% of the volume released by the supply system. Past research has indicated that water demand in Fortaleza is highly heterogeneous and that socioeconomic factors play an important role on consumption habits.

6.2.2 Water tariff structure

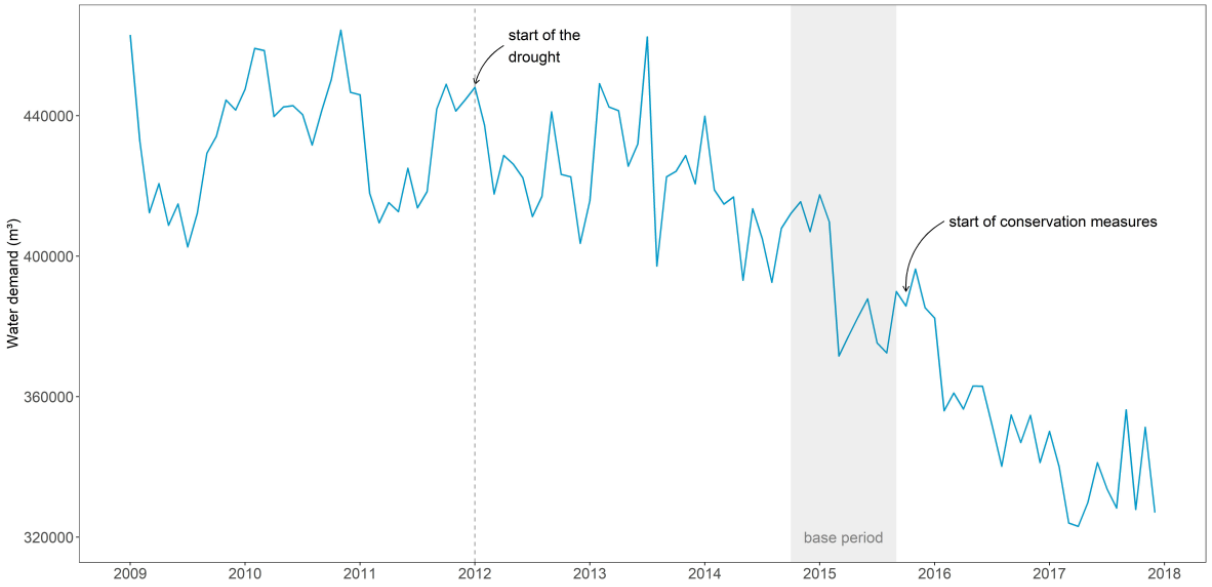
During the period between 2012 and 2018, the northeast of Brazil suffered from a historic drought that significantly impacted its economy and water storage (PONTES FILHO *et al.*, 2020). The main reservoirs of Fortaleza's supply system were affected by the 2012-2018 drought, resulting in a significant reduction in water availability. To encourage domestic water conservation, which accounts for more than 80% of Fortaleza's water demand, the local water company implemented a contingent tariff.

The contingent tariff was implemented in December 2015 (Figure 22) and defined a minimum reduction of 20% of the average consumption between October 2014 and September 2015. If a household did not meet this reduction goal, an extra charge of 110% on the exceeded volume would be added to the bill, i.e., the contingent tariff is calculated on the difference between the volume consumed and the goal. This percentage was updated to 120% in October

2016. Water price follows an increasing block tariff structure (Table 15); thus, the contingent tariff also varies with the consumption block of the household. Users with a monthly consumption of up to $10m^3$ did not have to pay the contingent tariff.

For example, for a household that had a mean consumption of $14m^3$ between October 2014 and September 2015, the goal was to use up to $11m^3$, corresponding to a 20% reduction in the monthly consumption. If in a certain month of 2017 the water demand of this household was $13m^3$, in addition to the water tariff $(13 * 4.51)$, they would have to pay the contingent tariff, which would be charged over the $2m^3$ that exceeded the consumption goal $(1.2 * 2 * 4.51)$. The base price here corresponds to the second block of consumption (*R*$4.51 in 2017).

Figure 22 – Total domestic water demand (m^3) in Fortaleza from 2009 to 2017. The baseline period was used by the local water company to calculate the reduction goal for each household.



Source: The author.

Although we consider these specific conditions in the prediction model, the methodology could be replicated under different price-associated water conservation measures.

Table 15 – Water tariff in Fortaleza for each consumption category for 2016 and 2017.

Monthly consumption (m^3)	2016 (BRL)	2017 (BRL)
0 to 10	2.79	3.48
11 to 15	3.61	4.51
16 to 20	3.92	4.88
21 to 50	6.71	8.36

Source: The author.

6.2.3 Predictive model

The predictive model has three explanatory variables: previous water demand, monthly seasonality of water demand, and the cost of the penalty fee, i.e., the contingent tariff cost per household. The model was tested for multiple leading times, ranging from one to twelve months.

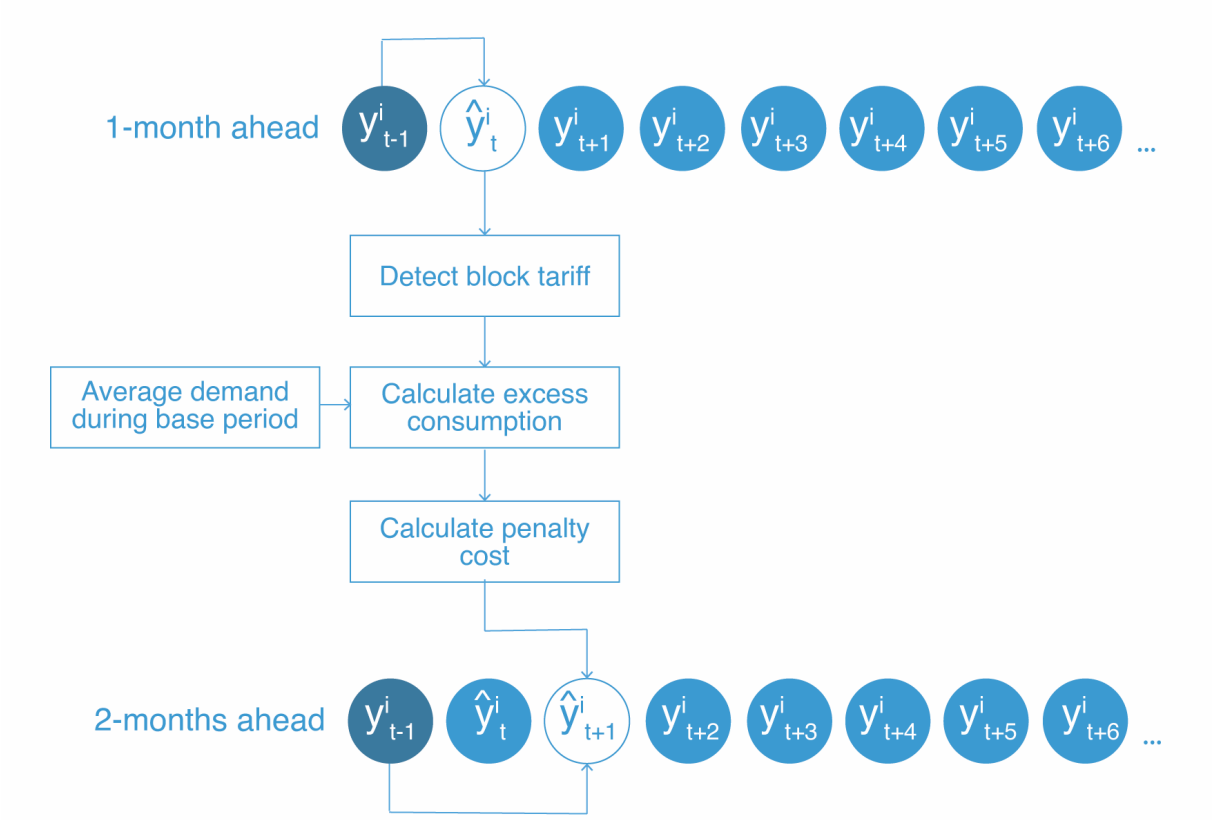
The penalty fee was calculated as the cost of the volume of water consumed in the previous month that exceeded a threshold. This threshold sets how much water should be saved and is a percentage of the average monthly water consumption of the household for a baseline period. Here, the baseline period goes from October 2014 to September 2015 and the threshold is 20%.

At each month, the predictions for the previous month are used to determine the tariff block of each household. Then, we calculate the volume of consumed water that exceeded the threshold and how much it cost for the user. For example, when calculating water consumption at n-months ahead, the predictions for the month $n-1$ are used to assess the water conservation measure (Figure 23). This strategy allowed us to avoid the simultaneity issue associated with water consumption modelling under block tariff policies.

Previous studies have used different price variables in econometric models of water demand, and there is not a generally accepted approach. Many authors find it more appropriate using the marginal price, i.e., the cost of increasing the water consumption at each time step (RINAUDO *et al.*, 2012), while others prefer the average price (ZHANG *et al.*, 2017) or both (MA *et al.*, 2014; DEYà-TORTELLA *et al.*, 2016). Although some researchers argue that the users might be more influenced by the average price (DEYà-TORTELLA *et al.*, 2016), in case of a contingent tariff policy, they might pay special attention to the additional charge expressed on the bill.

In addition to the lagged water consumption and the price component, a seasonal variable was included to account for seasonal behavior. This variable corresponded to the seasonal component extracted for each household with the Seasonal and Trend decomposition using Locally estimated scatterplot smoothing (STL) method. This approach captures different patterns of seasonal behavior and adds more information to the model than the usual approach of using 11 dummy variables for the months. We chose a machine learning regression model that has been widely used for electricity and wind prediction, Gradient boost regression. This algorithm also performs better than other linear and machine learning models in predicting

Figure 23 – The predictive model has an autoregressive component (previous month water demand) and the penalty fee as explanatory variables, in addition to the seasonality of the corresponding month. Starting from January, the water demand in December would be used to calculate the cost of the contingent tariff. For the next month, the penalty cost is calculated using the predicted water demand in January.



Source: The author.

residential water demand (LEE; DERRIBLE, 2020). The predictive model can be summarized in the following steps:

(i) Select a dataset χ of monthly household water demand and set a time horizon n (in months) for the predictive model.

(ii) Extract the seasonal component s_i of each household's water demand time series y_i using the STL method.

(iii) Set a consumption reduction goal or threshold and the penalty cost policy $p(.)$. The goal might be a percentage of the average consumption over a certain period, named the baseline consumption b_i.

(iv) Split the dataset into two subsets for training and validating the model. Initialize the gradient boosting model at month $t = 1$, setting the predictive variable y to $\hat{y}_t^i$ and the predictors to s_i, $p(y_{t-1}^i, b^i)$, and y_{t-1}^i. Choose arbitrary values for the main parameters of the model i.e., the number of trees, the minimum number of observations in each node and

the learning rate (usually ranges from 0.001 to 0.1).

(v) Run the model again using the predicted water demand $\hat{y}_t^i$ to calculate the penalty cost and estimate y at month $t+1$. If the water tariff follows an IBR structure, it might be necessary use a function $f(\hat{y}_t^i)$ to set the tariff block to the household prior to calculating the penalty cost. Repeat this procedure until $t = p$.

(vi) Compute model's performance $D(\hat{y}, y_t^i)$ on the training and testing sets and compare the measures to adjust the parameters and avoid overfitting the model.

The tabular version of the algorithm is described below:

Initialize: Set the variable $\hat{y}$ equal to y_t^i.

- Calculate the baseline consumption and the reduction goal

- Decompose the water demand time series using STL and extract its seasonal component s^i

repeat

- Determine the tariff block of each household based on the consumption of the previous month using a function $f(y_{t-1}^i)$. This step can be ignored if the water tariff does not follow an IBR structure.

- Calculate the penalty cost using a function $p(y_{t-1}^i, b^i)$

- Estimate a gradient boosting regression model that predicts y using s_i, $p(y_{t-1}^i, b^i)$, and y_{t-1}^i as predictors

- Compute model's performance using the selected measure(s) $D(\hat{y}, y_t^i)$ **until** $t = n$

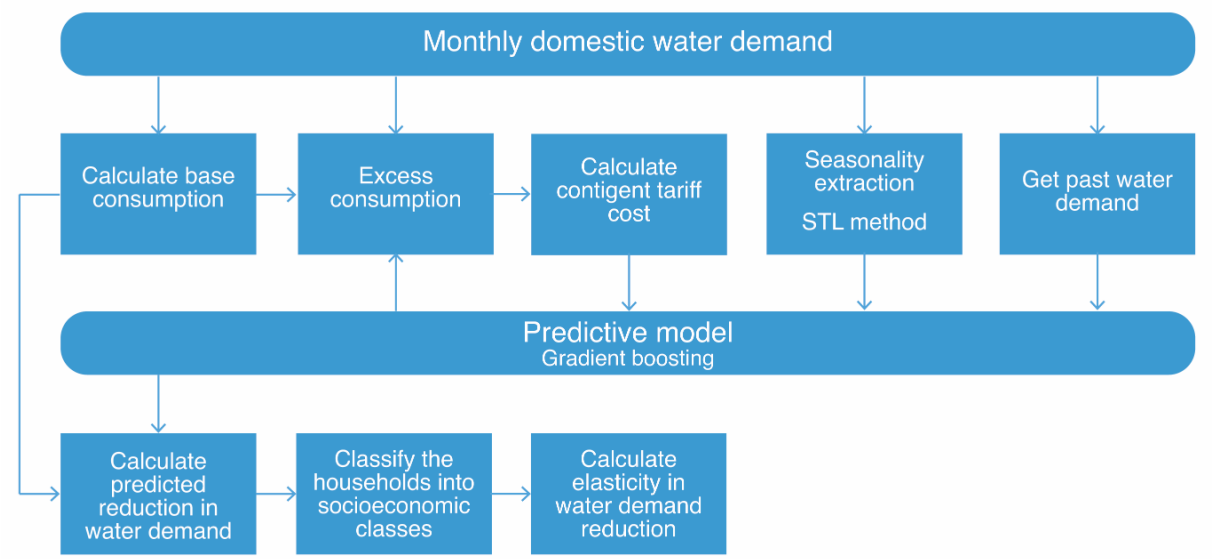
- Adjust model's parameters based on the performance of the training and testing subsets.

The model was validated with a classical out-of-sample evaluation and was trained for the year of 2016 and tested for the year of 2017. Figure 24 provides a general outline of the predictive model and the performed analysis.

6.2.4 Seasonality extraction

The water demand time series was decomposed into trend, seasonal and remainder components using the STL method (CLEVELAND *et al.*, 1990). This procedure was used to extract the seasonality of water consumption for each household. STL consists in sequential applications of the local regression model and provides an additive decomposition of the original

Figure 24 – Predictive model outline. The contingent tariff cost is recalculated as new predictions become available.



Source: The author.

signal (D) into three components:

$$D(t) = S(t) + T(t) + R(t)$$

where S, T and R are the seasonal, trend and remainder components, respectively. The algorithm work as follows:

The local regression smoothing estimates a function $g(x)$ for the independent variable at any value of x rather than for the measurements x^i of the dependent variable. To calculate the regression curve g, an initial value for the parameter q is chosen; q values of x^i that are closest to x are selected and weighted on their distance from x. For $q \leq n$, where n is the number of observations in the data set, the neighborhood weight for x^i is calculated as follows:

$$v_i(x) = W\left(\frac{|x_i - x|}{\lambda_q(x)}\right)$$

where $v_i(x)$ is the neighborhood weight for x_i, $\lambda_q(x)$ is the distance between x and the most distant x_i. For $q > n$, $\lambda_q(x)$ is multiplied by $\frac{q}{n}$. $W(u)$ is the tricube function, expressed as:

$$W(u) = \begin{cases} (1-u^3)^3, & \text{if } 0 \leq u < 1 \\ 0, & \text{if } u \geq 1 \end{cases}$$

Next, a polynomial of degree d is fit to the weighted data at (x_i, y_i). The value of d can be 0 (constant), 1 (locally linear) or 2 (locally quadratic). In this paper, $d = 1$. The fitted function corresponds to $g(x)$. It is possible to add a robustness weight ρ_i for each pair (x_i, y_i) by multiplying it by v_i.

STL consists of two nested loops (CLEVELAND *et al.*, 1990). In the outer loop, robustness weights are calculated for each time point. Initially, the trend and remainder component are set to 0 and ρ_i is set to 1. In the next loops, the remainder component is found by removing the trend and seasonal components calculated in the inner loop from the original series. The robustness weight is then calculated as follows:

$$\rho_i = B(\frac{|R|}{h})$$

$$h = 6 * median(|R|)$$

where B is the bi-square weight function, given as:

$$B(u) = \begin{cases} (1-u^2)^2, & \text{if } 0 \leq u < 1 \\ 0, & \text{if } u > 1 \end{cases}$$

The outer loop is repeated no times; if one does not wish to add robustness into STL, no should be set to 0. In this paper, no = 15. The inner loop follows these steps: (i) Detrend the original signal; (ii) Estimate a smoothing function using Loess for each cycle-subseries, where q is the cycle periodicity (e.g. for a monthly time series, q is set to 12) and d is equal to 1; (iii) Apply a low pass filter to the smoothed cycle-subseries, which consists in sequential applications of a moving average; (iv) Detrend the smoothed cycle-subseries; (v) Remove the seasonality from the series; (vi) Smooth the deseasonalized series using Loess. The STL decomposition can be easily performed using the stl function from base R.

6.3 Gradient boosting

Gradient boosting machines (GBM) (FRIEDMAN, 2001) is a learning method that converts weak learners, usually regression trees, into strong learners by combining them sequentially. The idea behind the method is that new weak learners can learn from the residuals of the output from the previous model; this ensemble technique is called bagging. For regression tasks, we want to find the function that best fits the data points in a set containing input variables x and a corresponding output variable y. To do this, the algorithm minimizes a loss function between y and the predicted values, in our case, the MSE.

$$L(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i^2)$$

The gradient boosting method consists in a combination of weak learners that are added together. The individual models f_m are added one after the other to improve model performance.

$$y_i = \sum_{m=1}^{M} f_m(x)$$

The weak learners, in this case, regression trees, are fitted on the residuals of the previous model. The general representation of GBM is expressed as follows:

$$F_m(x) = F_{m-1}(x) + \nu f_m(x)$$

meaning that the model f_m does not change the previously fitted model $F_{(m-1)}$. The term is a regularization parameter or the learning rate, which determines the number of iterations. Small values of the learning rate $(\nu 0.1)$ reduce the chances of overfitting. Gradient boosting applies a functional gradient descent method to minimize the loss function, where each new weak model is equivalent to the negative gradient of the MSE. The negative gradient is given as:

$$-g_m(x_i) = -[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}]_{Fx=F_{m-1}(x)}$$

The algorithm stops when the loss reaches a threshold, or the maximum number of trees is built. An important element to consider when fitting machine learning models (or any predictive model) is the bias-variance tradeoff and the chance of overfitting the model. If the algorithm misses important connections between the predictors and the response variable, the model will have a high bias, i.e., an elevated difference between predictions and the observed data. However, if during the model fits too perfectly to the training data, resulting in a high variance, it will not generalize well (overfit).

The best scenario when developing a model is to accurately capture the relationships between the variables during training but also make good predictions during training. In machine learning models, one can control the bias-variance tradeoff by controlling model parameters. The main parameters of GBM are the number of trees, which should not be too high to avoid overfitting; the minimum number of observations in each node, which defines how depth the tree might become; the learning rate or shrinkage, which relates to the size of the incremental steps, usually ranging from 0.01 to 0.1, and the distribution of the response variable, which in our case, was Gaussian.

In our framework, parameter tuning was performed in a trial and error manner, i.e., we defined arbitrary values for them, compared model performance for the train and test datasets and chose those parameters that resulted in comparable performances for both and could not be

improved anymore. The number of trees was set to 300, the learning rate to 0.1, and the number of observations per node to 10. All analyses were performed using R programming language. The gradient boosting model was implemented with the package gbm (GREENWELL *et al.*, 2020).

6.3.1 Performance assessment

Model performance was evaluated for the entire prediction horizon, i.e., for twelve months of the testing period. Two measures were used: RMSE and R^2.

$$RMSE_j = \sqrt{\sum_{i=1}^{n} (\frac{\hat{y}_{i,j} - y_{i,j}}{n})^2}$$

$$R_j^2 = \frac{\sum_{i=1}^{n} (y_{i,j} - \hat{y}_{i,j})^2}{\sum_{i=1}^{n} (y_{i,j} - \bar{y}_j)^2}$$

where yi, j is the observed water demand in household i at month j, yi, j is the predicted water demand in household i at month j, yi, j is the mean observed water demand at month j, and n is the number of households.

6.3.2 Elasticity of water demand reduction to price

Different scenarios of price increase were considered, based on the tariff for the previous year (2015 for the training and 2016 for the validation period): no increase, 5, 10, 15 and 25%. To calculate the elasticity of water demand reduction to price, we used the predictions for the year of 2016 obtained with the model. The reduction is related to the average consumption during the baseline period (October 2014 to September 2015).

$$E = \frac{\frac{\Delta R}{R}}{\frac{\Delta P}{P}}$$

where R is the monthly average reduction in water demand and P is the average water block tariff.

Water demand elasticity was assessed for different socioeconomic classes, as users' response to water conservation policies tend be heterogeneous. These classes were based on the criteria used by the IBGE, which is based on per capita family income. IBGE uses the minimum wage to classify the families in five classes (Table 16). The monthly per capita income of a household is divided by the minimum wage to find the correspondent socioeconomic class

(i.e., $Income = N * minimumwage$). We also compared the predicted monthly reduction with the actual reduction aggregated water demand.

Table 16 – Socioeconomic classes and number of households in each of them. The total number of household analyzed here is 37,689.

Class	Number of minimum wages	Number of households	Percentage of the total number of households
A	20 or more	53	0.14%
B	10 < N < 20	969	2.57%
C	4 < N < 10	5,186	13.76%
D	2 < N < 4	17,554	46.58%
E	N < 2	13,927	36.95%

Source: The author.

6.3.3 Public interest and media coverage

In this study, water demand reduction is associated with the implementation of a price control measure, which was expected to change public behavior. However, demand control policies may include other strategies, such as promotional events, water conservation education programs and mass media advertising campaigns (SHARMA; VAIRAVAMOORTHY, 2009). In Fortaleza, the water company created an app to encourage users to report leaks and frauds and promoted educational campaigns in schools, public buildings, and social media.

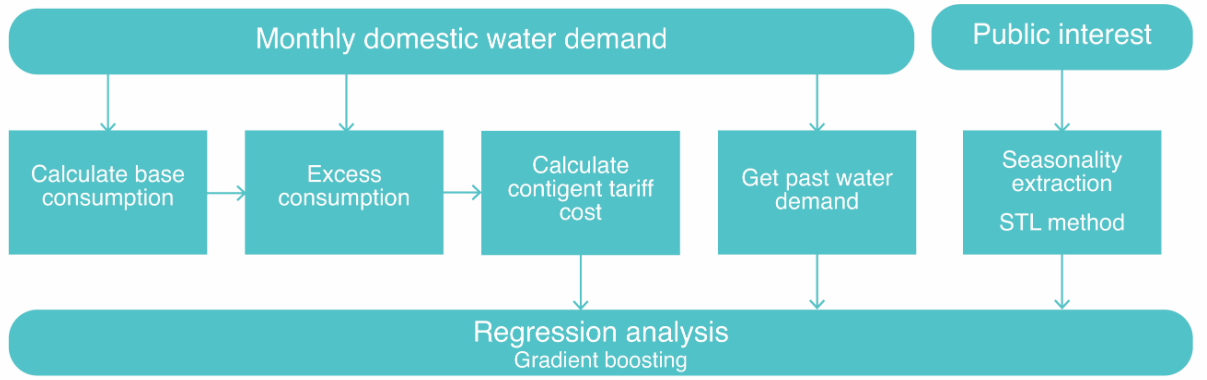
Google Trends data has been proven a useful tool for characterizing public response to certain matters and has been successfully applied to analyze private consumption (VOSEN; SCHMIDT, 2011) and to assess drought awareness (QUESNEL; AJAMI, 2017; KAM *et al.*, 2019). The idea here was to use the frequency of Google searches for the key words "contingent tariff" and "drought" to address people's interest on these matters and their awareness about the implementation of the tariff.

We acknowledge that mass media plays an important role on social systems (LUHMANN, 2000), hence media coverage on the contingent tariff might have influenced public response. For reference, we plotted the number of articles related to the contingent tariff published between 2012 and 2017, which were collected from the websites of the three main local newspapers (Tribuna do Ceará, OPovo and Diário do Nordeste). These sources have a strong online presence and usually share the news on social media such as Instagram and Twitter. Data was collected with web scraping using Python and the BeautifulSoup 4 library.

To assess the marginal response and the relative influence of public interest in drought and the contingent tariff on water demand, a regression analysis was performed using both as

explanatory variables (Figure 25). Water demand was predicted as a function of water demand in the previous month, public interest, and the contingent tariff cost for the previous month. Google search hits between 2012 and 2017 for the term “contingent tariff” and “drought” by users located in Fortaleza were used as a proxy for public interest in water scarcity, from which the trend component was extracted using the STL method.

Figure 25 – Regression analysis outline.



Source: The author.

The GBM algorithm was used to perform the regression. For this analysis, we used data from 2012 (beginning of the drought) to 2017. Note that here we fit the model using only observed data, i.e., the contingent cost is not iteratively calculated, since our intention was not to build a forecast but rather to assess the importance of the explanatory variables. For the same reason, seasonal water demand was not added as a predictor. The dataset was randomly split into 80% train and 20% test. After obtaining the regression model, we extracted the marginal response of each variable using partial dependence plots and their relative influence. The relative influence is measured with the reduction of squared error associated with each variable, i.e., how much worse the model’s performance would be without that variable.

6.3.4 Partial dependence plot

The PDP represents the marginal effect of independent variables on the response of a machine learning model (FRIEDMAN, 2001). The partial dependence of the response on a variable x_l is represented by:

$$\hat{f}_{x_l}(x_l) = E_{x_s}[\hat{f}(x_l, x_s)] = \hat{f}(x_l, x_s)P(x_s)dx_s$$

Where x_l is the independent variable analyzed in the partial dependence plot, x_s is the subset of the other input variables of the regression model f and $P(x_s)$ is the marginal probability density

of x_s. The function shows the effect of the variable x_l on the dependent variable by marginalizing over the other explanatory variables.

6.3.5 Data

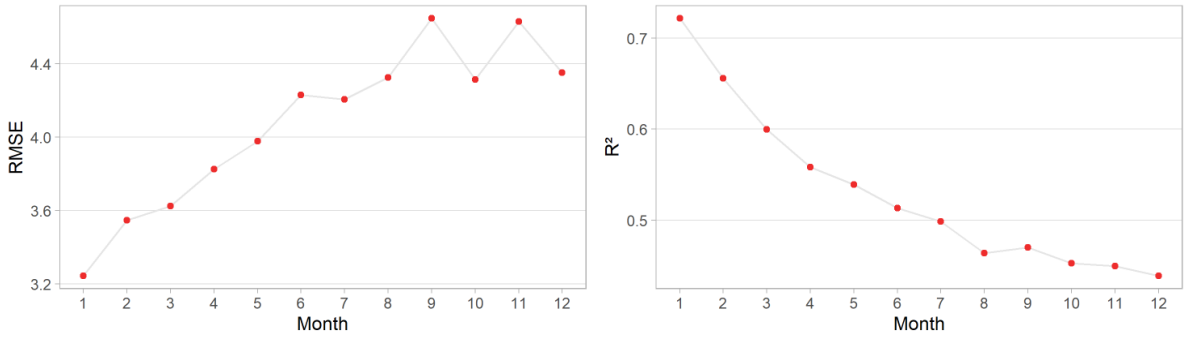
Monthly water demand data for the period between 2009 and 2017 from 45,141 households were provided by the CAGECE. This analysis focused on households with consumption up to $50m^3/month$. Households with monthly water consumption inferior to $1m^3$ per month or the ones in which the total water consumption between 2009 and 2017 was less than $5m^3$ were excluded from the dataset. The data cleaning process reduced the dataset to 37,689 observations.

Socioeconomic data from the 2010 Census were used to classify the households. Average per capita income is available at the census tract level, territorial units containing a maximum number of households that allow a survey to be carried out by a single person. Fortaleza is divided into 3,043 census tracts, and 2,586 of them are attended by CAGECE's water supply.

6.4 Results

Model performance was evaluated for each month of the testing period (Figure 26). The model presented reliable predictions in terms of RMSE and R^2 for a short-term horizon (1 to 6 months ahead), and satisfactory results for a medium-term horizon (7 to 12 months ahead). The autoregressive component was the most important, i.e., removing it from the model would mean a significant increase in the loss function. This suggests that water demand is strongly dependent on past use.

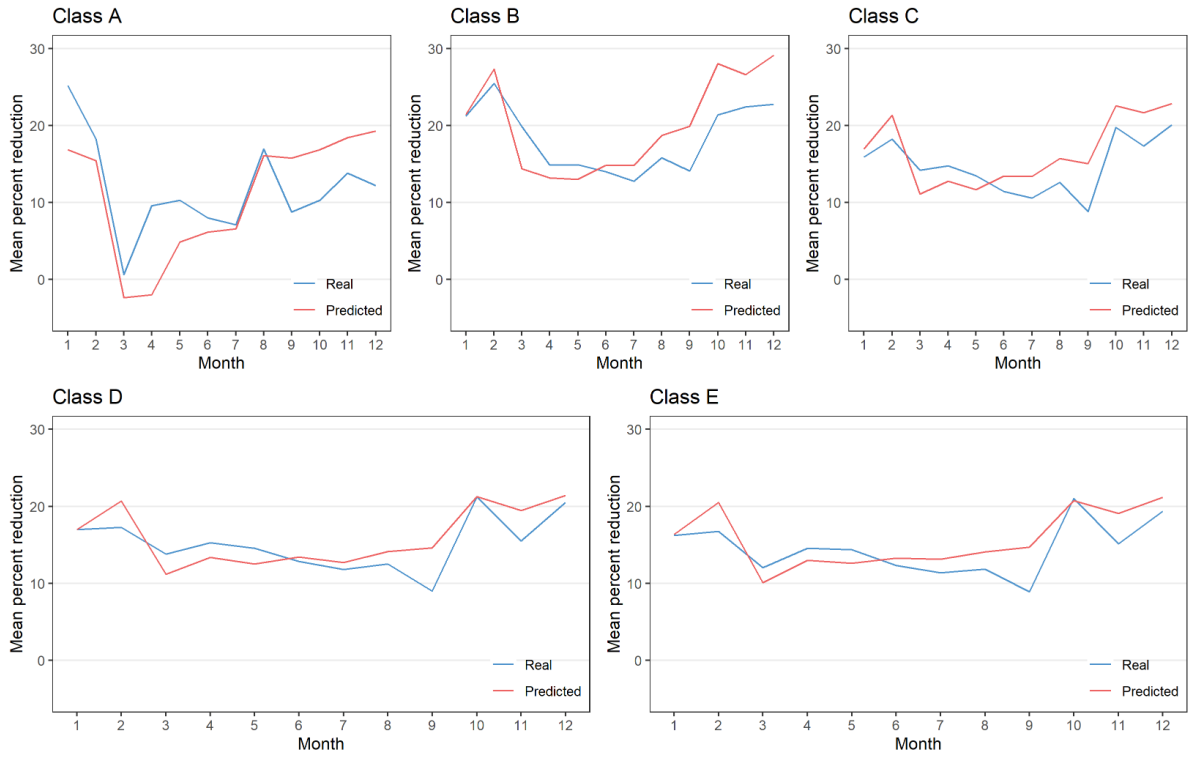
Figure 26 – Model performance.



Source: The author.

A comparison between the predicted and observed mean percent reduction in residential water demand shows that the model provided accurate predictions (Figure 27). For this analysis, households were grouped according to their socioeconomic class, to assess variation in model performance and mean percent reduction in water demand. Classes D and E presented a rather regular behavior during the year, with an average reduction of 14.73% and 13.99%, respectively. Households in class B had the largest reduction in water demand: 17.58% over the year. Class A, with the smallest reduction (11.22% on average), presented a peak in January but almost no change in March.

Figure 27 – Real and predicted monthly reduction in aggregated water demand for the year of 2017 for each socioeconomic class.

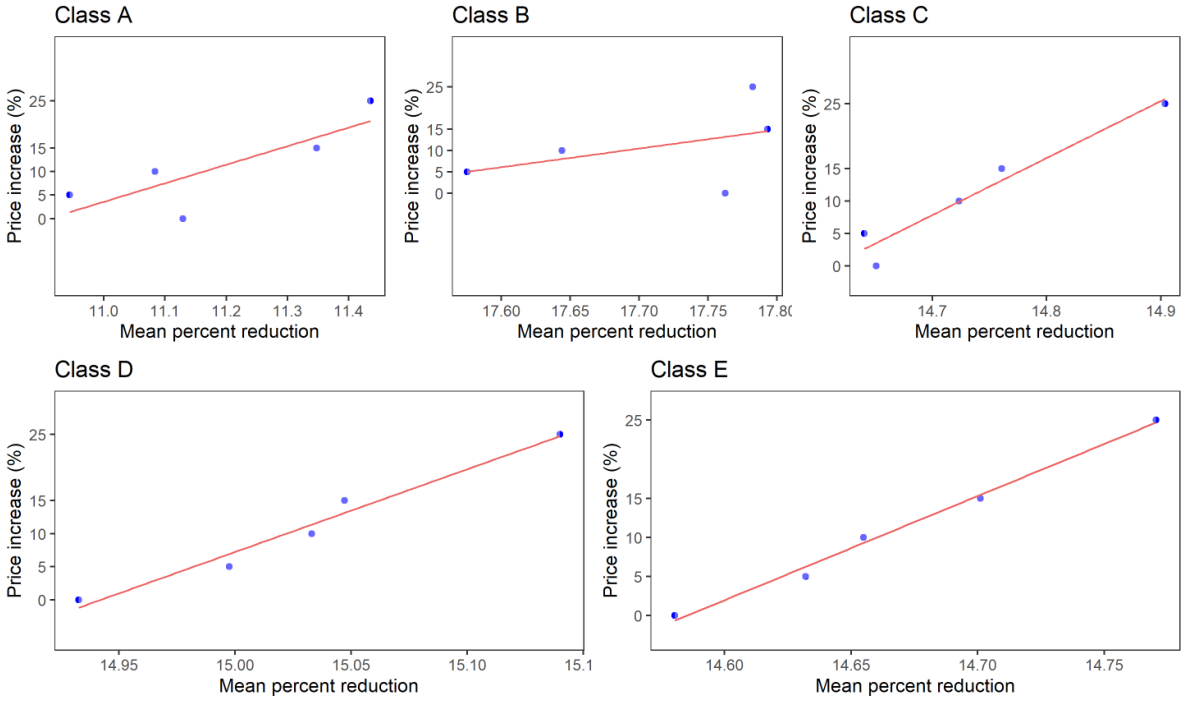


Source: The author.

The reduction in water demand was revealed inelastic to tariff variation (Figure 28). These results suggest that the contingent tariff itself would be enough to encourage a reduction in water consumption in all socioeconomic classes. However, the policy has adverse effects on each type of consumer. While the water tariff represents less than 1% of the average per capita income of classes A and B, it is about 23% of the income of class E, which represents 37% of the households (Table 17). The lower income classes had the lowest per capita consumptions during the baseline period, but still managed to reduce their demand after the implementation

of the contingent tariff. Except for households in class B, none of the classes would reach the 20% reduction goal. Class B also had the highest average daily per capita consumption (Table 3) during the baseline period.

Figure 28 – Elasticity of water demand reduction to price for each socioeconomic class.



Source: The author.

These findings agree with other studies that also found water demand is inelastic to price variation (RINAUDO *et al.*, 2012; DEYà-TORTELLA *et al.*, 2016). Also, Zhang *et al.* (2017) showed that increasing block policies are not effective to encourage a reduction in water consumption. Ma *et al.* (2014) indicated that the highest income group is not sensitive to price changes, while residents from the lower income group respond to marginal price and might even compare the tariff for different blocks to optimize their benefit. André and Carvalho (2014) found similar values of water demand elasticity to price in Fortaleza using survey collected data. The advantage here is that we used only secondary data to calculate elasticity for different socioeconomic classes.

Overall, the results indicate that the restriction policy might be unfair with the lower income classes, for which the tariff represents a significant percentage of their income and still enforced a reduction in its already low daily per capita demand. As stated by Bernoulli (1954), benefit perception depends on the individual perception of cost. Hence, a small increase in water cost has a more significant effect on the economic value attributed to water for lower income

classes.

In a scenario where the customers must pay an additional charge for their excess consumption, price increase does not seem to affect consumer behavior. This result can be explained by the fact that the customers might be at the kink point of the block rate schedule or their willingness to pay for water rises under drought conditions, since it represents only a small percentage of their income. The first is the most reasonable explanation for classes D and E, while the second is consistent with higher income classes. Another aspect to be considered is the reservation capacity of households (water tanks or cisterns, private borehole drilling), which is higher for wealthy customers (GRANDE *et al.*, 2016), who might be able to maintain their standards and still reduce the water volume from public supply.

Table 17 – Reduction in water demand elasticity to price increase and characteristics of the socioeconomic classes.

Socioeconomic class	A	B	C	D	E
Elasticity of water demand reduction to price	0.515	0.212	0.426	0.314	0.295
Number of households	53	969	5,186	17,554	13,927
Percentage of the average per capita income related to the water tariff (%)	0.46	0.89	2.00	3.77	22.97
Average daily per capita consumption (L/hab/day) for the baseline period	102.25	123.36	105.19	96.78	94.96
Average daily per capita consumption (L/hab/day) after the restriction measures	90.06	104.99	92.37	84.60	83.69
R^2	0.69	0.69	0.75	0.74	0.73
RMSE	4.37	4.08	3.26	3.01	3.05

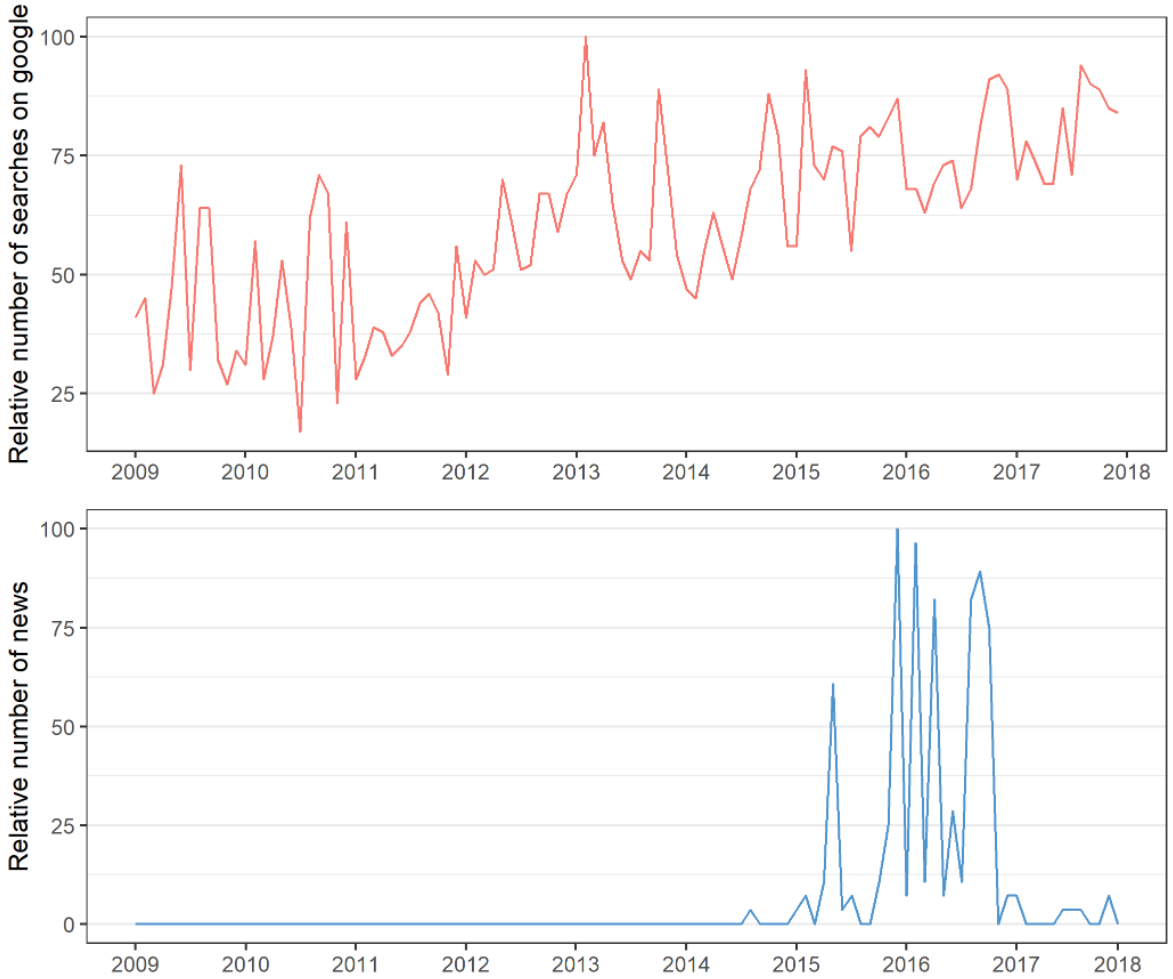
Source: The author.

It is important to bear in mind that the consumers are not necessarily aware of the pricing policy structure. Although the contingent tariff is clearly expressed on the water bill, increasing block tariff scheme is not detailed for households.

A clear increasing trend in public interest is observed after 2012 (when the drought started), while the number of news related to the restriction measure peaked in 2016 (Figure 29). While this could imply that the public was well informed about pricing policy, the finding cannot be extrapolated to all customers, since not all households have access to internet.

A regression analysis between water demand, public interest, the contingent tariff, and past water demand was performed for each socioeconomic class (Table 18). The relative importance values imply that an increase in the cost associated with the contingent tariff has a higher influence on consumer behavior than information on drought. Also, it seems that residents

Figure 29 – Public interest and media coverage on the contingent tariff policy.



Source: The author.

with higher income have a more significant response to both the contingent tariff and information on drought compared to residents in classes with lower income.

Table 18 – Relative importance of the explanatory variables of the regression model between water demand, past water demand, public interest, and contingent tariff.

Class	A	B	C	D	E
Past water demand	85.87	95.78	98.31	98.67	98.55
Contingent tariff cost	10.17	3.90	1.59	1.29	1.42
Seasonal public interest	3.96	0.32	0.10	0.03	0.02
R^2	0.69	0.69	0.75	0.74	0.73
RMSE	4.37	4.08	3.26	3.01	3.05

Source: The author.

PDPs were plotted for each regression model (Figure 30). The results indicate that water cost has an inverse relationship with water demand for all households, while an increase in

the interest in the drought has little effect on consumer habits. It is worth mentioning that class A is the only one to present a direct relationship between public interest in the drought and water demand. However, we should be careful when interpreting these results since class A has a low number of households.

6.5 Conclusion

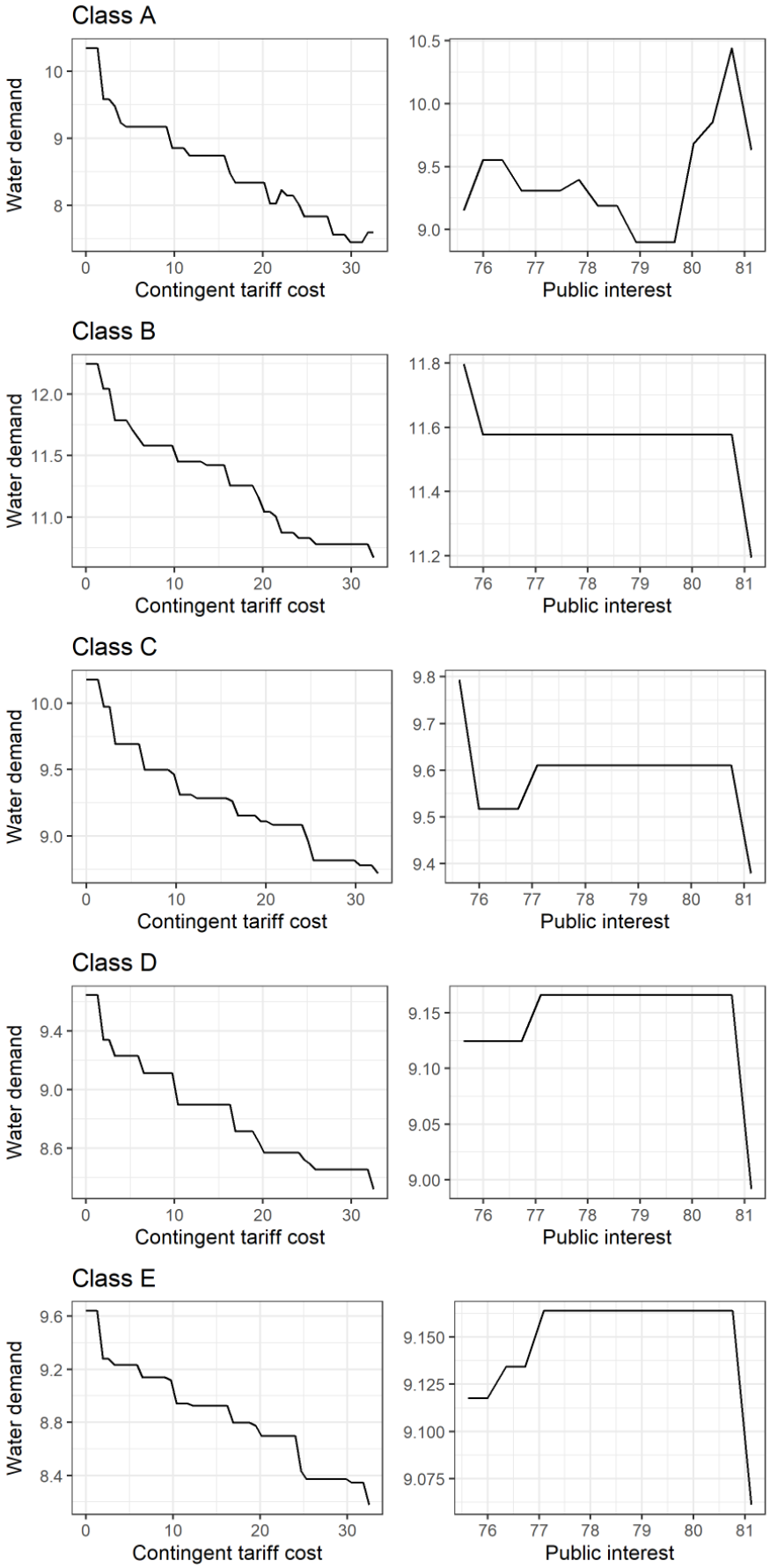
The main objective of this research was to address the influence of a contingent tariff on a predictive model of water demand in Fortaleza, Brazil. The model contained an autoregressive component and variables assessing seasonality and the cost associated with the contingent tariff. This study has found that the contingent tariff was effective and resulted in a 11-17% reduction in residential water demand. Also, reduction in consumption was inelastic to price increase in all socioeconomic classes.

The evidence from this study suggests that a price policy that associates with a contingent tariff could be unfair to lower income households, for which the tariff represents a large percentage of household income. Hence, although the strategy warrants a high revenue for the water company (that can be allocated to water security projects), its equity is questionable. Managers should be careful when implementing pricing policies to ensure the affordability of water services to all consumers.

The findings of this study imply that price-related water demand control policies are effective, while drought awareness is less likely to encourage consumers to save water. The increase in public interest in the drought does not necessarily indicate that consumers are well informed about the risks associated with it. It is crucial that the users are aware of the water resources management strategies and the implications of their habits rather than having a limited perception of drought. This can only be accomplished if social dynamics aspects are considered when designing drought plans and policies.

The framework proposed here is flexible and can be useful for water companies planning to implement price-related measures to encourage water demand reduction. The predictions at the household level can be useful to design policies for different classes of consumers. The predictive model can be used to verify at what extent the changes in the price policy could influence water demand.

Figure 30 – Partial dependence plots for public interest and the contingent tariff cost. A regression model was built for each socioeconomic class. Public interest is dimensionless.



Source: The author.

7 WATER INFRASTRUCTURE PLANNING UNDER CLIMATE VARIABILITY

> Eu atravesso as coisas - e no meio da travessia não vejo! - só estava mesmo era entretido na ideia dos lugares de saída e de chegada. A gente quer passar um rio a nado, e passa; mas vai dar na outra banda é num ponto mais em baixo, bem diverso do em que primeiro se pensou. (ROSA, 2019)

7.1 Introduction

Managing water under uncertainty and risk in urban systems is challenging, especially when water demand and urbanization tend to increase. Climate change and spatial and temporal variability of precipitation and temperature might change the intensity and frequency of extreme events, directly impacting water availability.

Long-term planning of water supply investment and short-term management decisions comprise a potentially large number of options that are difficult to tackle, especially in an uncertain environment (TRINDADE *et al.*, 2019). Capacity expansion of water systems includes the implementation of new water sources or the improvement of the existing infrastructure to meet growing demand. This problem formulates as a multi-stage model and involves decisions related to how much and when to invest in different water supply facilities at minimal cost (FRAGA *et al.*, 2017).

The capacity expansion optimization is a large-scale problem with a complex solution, that is usually solved approached with dynamic programming, stochastic dynamic programming, or multi-objective optimization (XIONG *et al.*, 2018). These strategies have a high computational cost (MORTAZAVI-NAEINI *et al.*, 2014), and are not suitable when considering multiple water sources in a long planning horizon. For this reason, the studies found in the literature do not offer flexible models that are also able to integrate operational and capacity expansion decisions of water supply.

To address this issue, we (i) formulate the capacity expansion problem as an optimization model and solve it with Stochastic Dual Dynamic Programming (SDDP) (PEREIRA; PINTO, 1991) and (ii) extract the operational rules obtained from the model using machine learning. SDDP was developed to overcome the curse of dimensionality (i.e. the increase in computational cost due to the exponential increase in stage dimension) and is one of the only

techniques available that can explicitly consider uncertainties (ROUGé; TILMANT, 2016). This technique is widely used to solve problems of operation of complex hydroelectric systems but has not yet been applied to the problem of capacity expansion of water supply. SDDP approximates the cost function with piecewise linear functions, avoiding the need to list all possible combinations of the capacity of the water sources under consideration. The integration of machine learning with the stochastic optimization algorithm results in a powerful planning tool, with the advantage of being adaptive, flexible and at the same time able to explicitly incorporate uncertainties.

7.2 Methodology

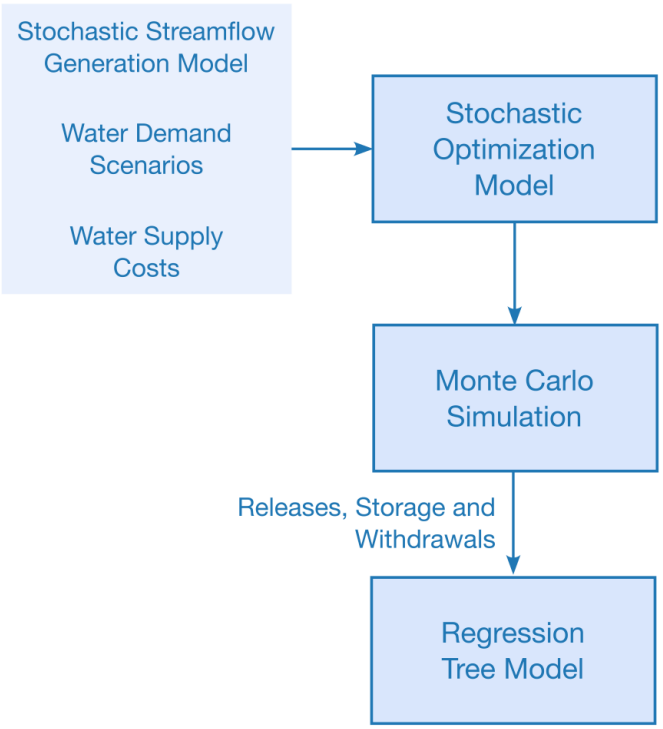
The problem of capacity expansion and operation of a water supply system was formulated as a stochastic optimization model. The model provides annual expansion decisions and monthly operational rules for a 30-year planning horizon, considering both conventional and unconventional water sources and uncertainties in water availability. The model minimizes the costs of expansion, operation, and maintenance of the water infrastructure for a tolerable risk of supply failure, using SDDP. After obtaining the optimal expansion policy, we simulate it using a Monte Carlo sampling scheme, i.e., we perform several simulations of the policy with random sampling of the reservoir inflow series. For each simulation, we record operational decisions (releases, reservoir storage and withdrawals from alternative water sources) and feed them to a regression tree model. In this final step, we obtain general reservoir operating rules to guide the water allocation process and decision making of the water system stakeholders. The methodological strategy, based on (LABADIE, 2004), is summarized in 31.

7.2.1 Case study

The optimization framework was applied to a case study for the RMF, Brazil. The region is supplied by eight storage reservoirs, pump stations and canals that transfer water from the Jaguaribe River basin, through the JMS. Five of them supply the Metropolitana basin, 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.

The expansion of the RMF water supply system (Figure 32) involves the inclusion of some alternative water sources, including: (i) wastewater reuse (destinated exclusively for

Figure 31 – Methodological strategy.



Source: The author.

industrial use), (ii) a desalination plant, to be installed on Iracema beach, and (iii) transbasin diversion through the PISF.

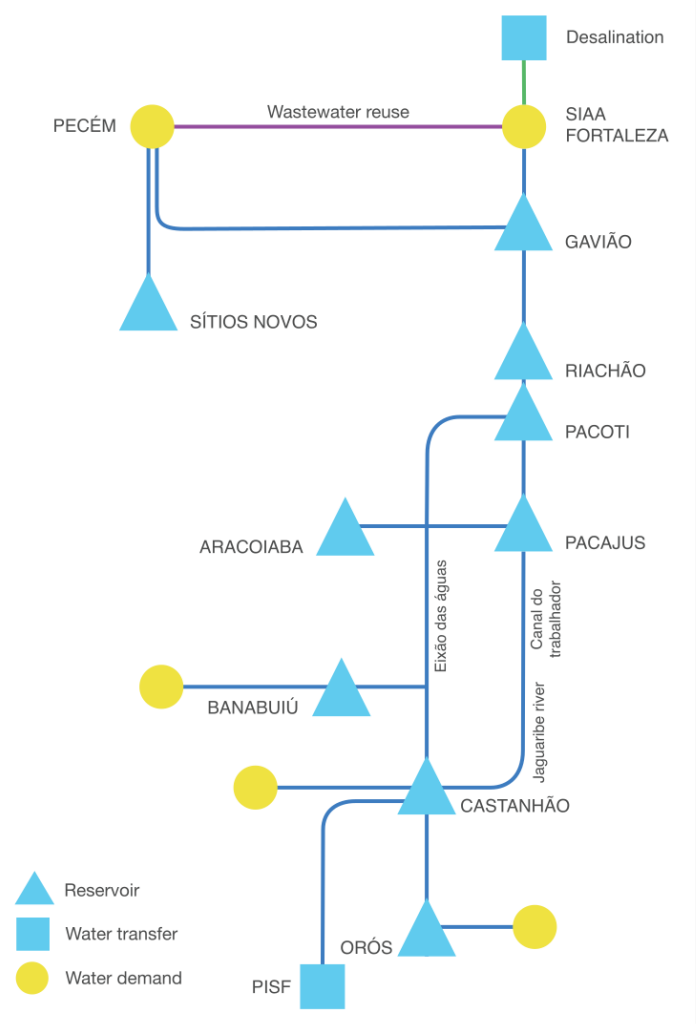
7.2.2 Water supply sources

Fortaleza's Integrated Water Supply System (IWSS) relies solely on the JMS, comprising eight surface reservoirs (Orós, Castanhão, Banabuiú, Aracoiaba, Pacajus, Pacoti-Riachão, Sítios Novos and Gavião). The initial (30% of maximum volume) and maximum volume considered in the model are described in Table 19.

100 synthetic streamflow series were generated for the each reservoir. In the simulations, we consider that all series have the same probability of occurrence ($p = 1/100$). The streamflow series generation methodology combines a generalized linear model (GLM), for determining the temporal structure, with copulas, for modeling the joint distributions of spatial dependence. The model is more efficient at replicating long-term droughts than the classic autoregressive and moving average (ARMA) model. The method is described in detail in (PORTO *et al.*, 2021).

The desalination plant is expected to have a minimum flow of 0.2 m^3/s, and it can

Figure 32 – Single-line diagram of the JMS.



Source: The author.

Table 19 – Initial and maximum volume of the surface reservoirs considered in the optimization model.

Reservoir	Initial volume (hm^3)	Maximum volume (hm^3)
Banabuiú	480.3	1601
Orós	582.0	1940
Pacajus	72.0	240
Pacoti	114.0	380
Gavião	9.9	32.9
Riachão	14.1	46.95
Sítios Novos	37.8	126
Castanhão	2010.0	6700
Aracoiaba	51.2	170.7

Source: The author.

deliver 1 m^3/s by the end of the planning horizon (30 years). The desalination plant needs at least three years to be installed, hence, it can only supply water from the fourth year in the planning horizon (this condition was added as a constraint to the optimization model).

Information regarding the implementation of wastewater reuse, such as the maximum flows and their implementation stages, were extracted from the report "Feasibility study for two tertiary sewage treatment stations and for an automation and control system for water and sewage for CAGECE" (CAGECE, 2017). The project foresees the installation of tertiary sewage treatment technologies in two drainage basins of Fortaleza, and reclaimed wastewater will supply industrial demand. The project should take at least five years to be finished, hence this source is only available from the sixth year in the planning horizon.

Water transferred from the São Francisco river through PISF is delivered to the Castanhão reservoir; therefore, in the model equations, the source is considered as an additional inflow to this reservoir. The transposition of the São Francisco river started in 2021. In that year, 64.9 hm^3 of water from the PISF were transferred to Ceará - from this volume, 54.9 hm^3 were transferred to the RMF.

It is worth mentioning that there are still no studies on water transfer losses (and inflows) to the Castanhão reservoir, which may occur (i) by infiltration in the distribution network and/or (ii) due to evaporation from the small reservoirs located in the stretch. The losses – which were not considered in the model, as they have not yet been properly estimated – can reduce the water availability expected by the transfer and increase the need to use alternative water sources, such as desalination and the reuse of effluents.

Table 20 – Minimum and maximum capacity of the water sources included in the optimization model.

Water Source	Minimum withdrawal	Maximum capacity (m^3/s)
Desalination	20% of the installed capacity	1
Wastewater reuse	0	4.5
PISF	0	10

Source: The author.

The minimum withdrawals and the maximum capacity of all sources considered in the model are described in Table 20.

7.2.2.1 Water demand

We estimate the urban supply, agriculture and industrial demands for each reservoir using a linear growth function (Table 21). The water demand in the first year of the planning

horizon corresponds to the demand in 2020, obtained from the COGERH.

Table 21 – Water demand projections.

Reservoir	Urban supply demand (m^3)		Agriculture demand (m^3)	
	Year 1	Year 30	Year 1	Year 30
Castanhão	0.61	0.73	11.93	13.87
Banabuiú	0.07	0.09	0.92	1.04
Orós	0.23	0.28	3.43	3.73
Aracoiaba	0	0	0	0
Sítios Novos	1.2	3	0	0
Pacoti	0	0	0	0
Riachão	0	0	0	0
Gavião	9.27	18.33	0	0
Pacajus	0	0	0	0

Source: The author.

We considered a percentage of water distribution losses in the of 45% (27% physical losses and 18% apparent losses), with a 20% reduction by the end of the planning horizon. Thus, in year 30, the percentage of losses is 36%. These losses were incorporated into the Gavião reservoir demand, which supplies the (increase in projected demands).

7.2.3 Water supply costs

Installation, transfer, operational and maintenance costs were obtained from a systematic review of the literature and information from water utility companies. All costs were converted into net present value (Table 22). For a detailed description of the costs assessment, see Ribeiro *et al.* (2022).

Table 22 – Investment, operational and maintenance (O&M) costs of the water sources included in the optimization model.

Reservoir	Fixed O&M cost ($R\$/m^3$)	Variable O&M cost ($R\$/m^3$)	Investment cost ($R\$/m^3$)
Desalination	1.01	1.80	1.12
Wastewater reuse	0.60	*0.83	1.83
PISF	0.69	0.51	0

Source: The author.

In addition to the costs of operation and maintenance and installation of water sources, the optimization model also incorporates the supply failure cost (penalty). This cost was quantified as a penalization parameter that is multiplied by the deficit in water demand. As urban supply should be prioritized over agricultural water use (as stated by the state water policy (CEARá, 1992)), failing in supplying these demands results in different penalties (β_1 and β_2,

respectively). We also included a penalty associated with reducing the volume stored in the reservoirs below 20% of their maximum capacity (β_3). These costs were experimentally set to $\beta_1 = 8\ R\$/hm^3/month$, $\beta_2 = 6\ R\$/hm^3/month$ and $\beta_3 = 10\ R\$/hm^3/month$.

7.2.4 Optimization model

The decision variables are the monthly withdrawal from each source to meet the water demand, the monthly transfer between basins and the annual capacity expansion of each water source. The state variables are the volumes stored in the reservoirs and the installed capacities of each source. Except for desalination, water availability of all the other sources is directly or indirectly conditioned to climate variability.

Let q_t be the vector of inflows during period t, x_t be the vector of volume in storage at the beginning of time period t, f_t be the cost of system operation during period t, q_t be the cost of expanding the system during period t, and v be a terminal value function. The expected costs to be minimized from system operation and expansion from period 1 to period T are:

$$z = E[\sum_{t=1}^{T} f_t(x_t, q_t, u_t) + q_t(x_t, q_t, y_t) + v(x_{T+1}, q_T)]$$

The problem is optimized under a set of hydrological, physical, and institutional constraints. Hydrologic uncertainty associated with surface reservoir inflow was assessed with an equiprobable model of inflows obtained from the historical series. Reservoir operation was included using the water balance continuity equation, which provides monthly yield, evaporation, and spill.

$$s_{t+1} = s_t + q_t + C(r_t) - l_t - e_t(s_t) - x_t^{hum} - x_t^{agr} - i_t$$

Where s_t is the vector of storage volume at the beginning of time period t, q_t is the vector of inflows, l_t is the vector of spills, C is the system connectivity matrix, e_t is the evaporation loss, x_t^{hum} is the vector of urban supply withdrawals and x_t^{agr} is the vector of agriculture withdrawals. The indexes of the alternative water sources are 1 (wastewater reuse), 2 (desalination) and 3 (PISF). Evaporation was as a linear function of reservoir volume, as considering a nonlinear function would add a nonconvexity to the optimization problem. This assumption was considered valid as the percentage difference between evaporation loss calculated

with linear and nonlinear functions was less than 10% for all reservoirs. Monthly operation decisions are conditioned to the maximum capacity of the alternative water source ($Cap_{m,i}$) defined for the corresponding month.

$$x_{m,t} \leq Cap_{m,t}$$

Except for surface reservoirs, the installed capacity of all water sources can be monthly increased. $y_{t,i+1}$ represents the increase in the capacity of the water source m.

$$Cap_{m,t+1} = Cap_{m,t} + y_{m,t+1}$$

The installed capacity of the water sources can not be reduced in a further month.

$$Cap_{m,t+1} \geq Cap_{m,t}$$

Monthly withdrawals can not be greater than the water volume stored in the surface reservoirs.

$$x_{r,t}^{hum} + x_{r,t}^{agr} <= s_{r,t}$$

The unmet water demand will be multiplied by a penalty factor in the cost function.

$$u_{hum} >= d_t^{hum} - x_{r,t}^{hum} - x_{m,t}^{hum}$$

Where d_t^{hum} is the vector of the urban water demand in t, $x_{r,t}^{hum}$ is the withdrawal from surface reservoirs and $x_{m,t}^{hum}$ from alternative water sources and u_{hum} is the unmet urban water demand. Agriculture water demand has a different penalty factor, as urban supply is prioritized over agriculture use.

$$u_{agr} >= d_t^{agr} - x_{r,t}^{agr} - x_{m,t}^{agr}$$

Where d_t^{agr} is the vector of the agriculture water demand in t, $x_{r,t}^{agr}$ is the withdrawal from surface reservoirs and $x_{m,t}^{agr}$ from alternative water sources and u_{agr} is the unmet agriculture water demand. Once installed, the desalination plant must produce at least 20% of its total capacity.

$$x_{2,i+1} \geq 0.2 * y_{2,i}$$

Wastewater reuse is limited to a percentage (α) of the water consumed in the RMF, i.e. water from desalination and the Gavião reservoir (index = 9). (α) was set to 0.8.

$$x_{1,t} <= (x_{2,t} + x_{1,t}^{res=9}) * \alpha$$

The volume stored in the reservoirs should not be less than 20% of their maximum capacity.

$$u_{res} \geq (0.2 * s_{r,max}) - s_{r,t}$$

Where $s_{r,max}$ is the maximum storage of reservoir r and $s_{r,t}$ is the withdrawal from the corresponding reservoir.

The objective is to minimize the installation, operation and maintenance costs and the failure to supply water demand, while at the same time, maintaining the reservoir volume in at least 20% of its maximum capacity.

$$\sum_{m=1}^{3}(OMC_{fix,m} * Cap_m) + \sum_{m=1}^{3}(OMC_{var,m} * x_m) + \sum_{m=1}^{3}(IC_m * Cap_m) + \sum_{r=1}^{9}(TC_r * Cap_m) +$$

$$+\beta_1 * u_{hum} + \beta_2 * u_{agr} + \beta_3 * u_{res}$$

Where $OMC_{fix,m}$ is the fixed OM cost of the source m, $OMC_{var,m}$ is the variable OM cost of the source m, IC_m is the installation cost of the source m, TC_r is the cost of transferring water from reservoir r to another reservoir.

The optimization model was developed using the programming language Julia 1.8.5 (BEZANSON *et al.*, 2017) and the SDDP package (DOWSON; KAPELEVICH, 2021).

7.2.5 Risk Assessment

To calculate the risk associated with the optimal expansion strategy of the water supply system, we calculate the water supply failure associated with each reservoir. The water supply failure corresponds to the percentage of months in which the unmet fraction of the total demand is greater than 10%. Then, we calculate the frequency that the system attends water demand (reliability, as defined by Hashimoto *et al.* (1982)) and the magnitude of failure (monthly supply deficit).

7.2.6 Extraction of operating rules

After simulating the optimal policies obtained with the optimization model, we extracted the reservoir operation rule using a decision tree model. 100 simulations of the model were performed using the Monte Carlo method. During a Monte Carlo simulation, streamflow series are randomly sampled from the input probability distributions. Each sample set is an

iteration. The Monte Carlo simulation does this hundreds of times, and the result is a probability distribution of possible outcomes.

The operating rules were extracted using a decision tree model. A decision tree provides a set of rules for expressing the relationship between explanatory and response variables, which are represented with a tree structure (KRZYWINSKI; ALTMAN, 2017). The leaves represent class labels (classification) or estimates of the response variable (regression) and the branches or internal nodes represent the values of the tested variable. This strategy has been used by other researchers to derive operating rules (WEI; HSU, 2008; YANG *et al.*, 2016).

The inputs of the predictive model are monthly releases (to supply either urban or agricultural demands), reservoir storage and withdrawals from alternative sources. For each reservoir, we fit several decision trees; considering (i) the month in the year and (ii) whether the alternative water sources have been installed or not. The response variable is the release for both urban and agricultural supply (regression) or the decision to use or not water from wastewater reuse and desalination (classification). Data was split into training (80%) and test (20%) and the complexity parameter was tuned with the grid search method (we varied it from 0 to 0.4, by 0.01 increments). We used a repeated 10 fold cross validation setting with the accuracy (classification) and the R^2 (regression) as performance metrics. The minimum number of observations in a node to have a split was set to 20, while the maximum tree depth was set to 10.

The goal of this analysis is not to estimate an accurate forecasting model, but to extract information regarding the operating strategies indicated by the optimization model. Therefore, although the decision tree model might not the most suitable for making accurate predictions – RF (BREIMAN, 2001) and GBM usually have better performances (YANG *et al.*, 2016) – it is an excellent tool for inference and decision making.

7.3 Results

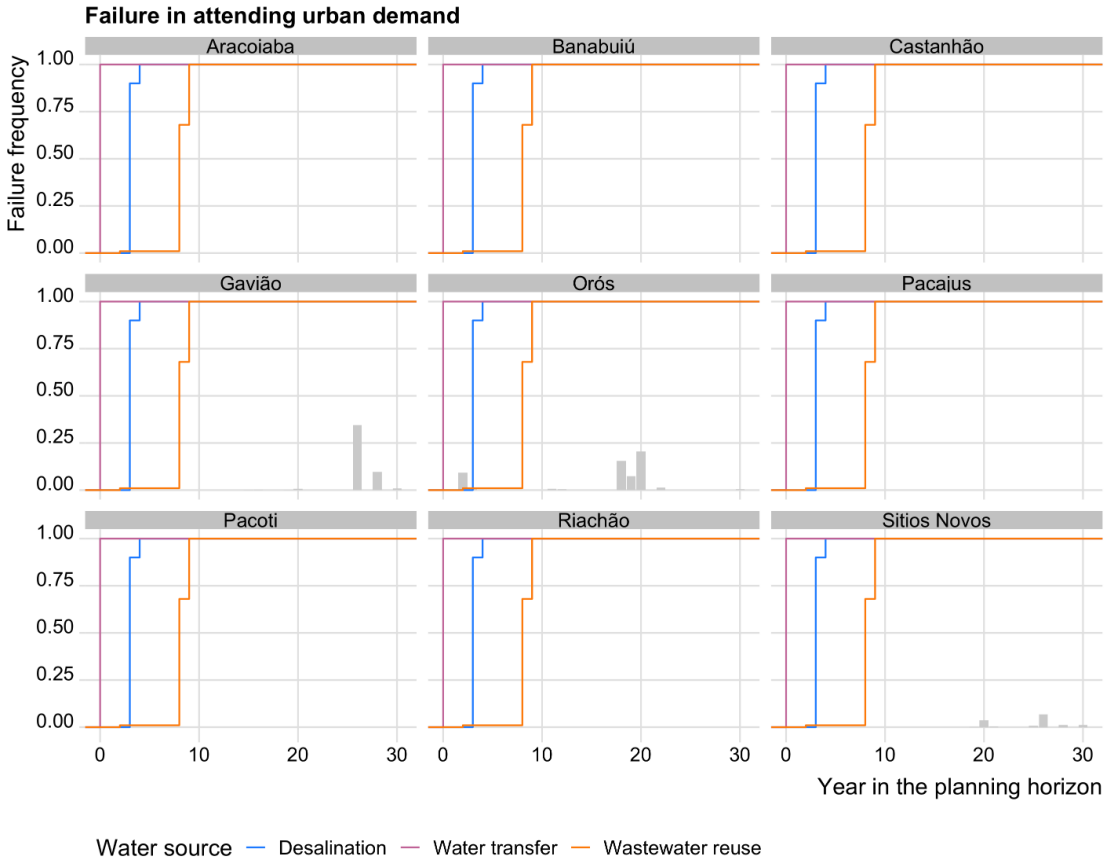
7.3.1 Optimal expansion strategy and Risk assessment

In all of the simulations, all three alternative water sources (i.e. water transfer from PISF, desalination and wastewater reuse) were needed to supply water demands. However, they were implemented in different time frames. We calculated the ECDF of the activation year of each alternative water source in all simulations to identify the most appropriate timing for installing them. While wastewater reuse should be necessary around the 8th year of the planning

horizon; desalination should be needed around year 3, and the water transfer from PISF from the first year in the planning horizon. We also computed the monthly water shortage (i.e. the difference between urban/agricultural water demand and release for these uses) per reservoir and the frequency of water supply failure (when the deficit is above 10% of the demand).

Figure 33 indicates that the failure frequency in attending urban demand is below 50% for all reservoirs during the planning horizon and except for Orós, it only happens in the last 10 years. The maximum average deficit is around 1 m^3/s in the 26th year of the planning horizon (Figure 34); which indicates that the growing demand of the RMF (supplied by Gavião reservoir) might result in reduced water security over the years or when the system is under severe drought conditions.

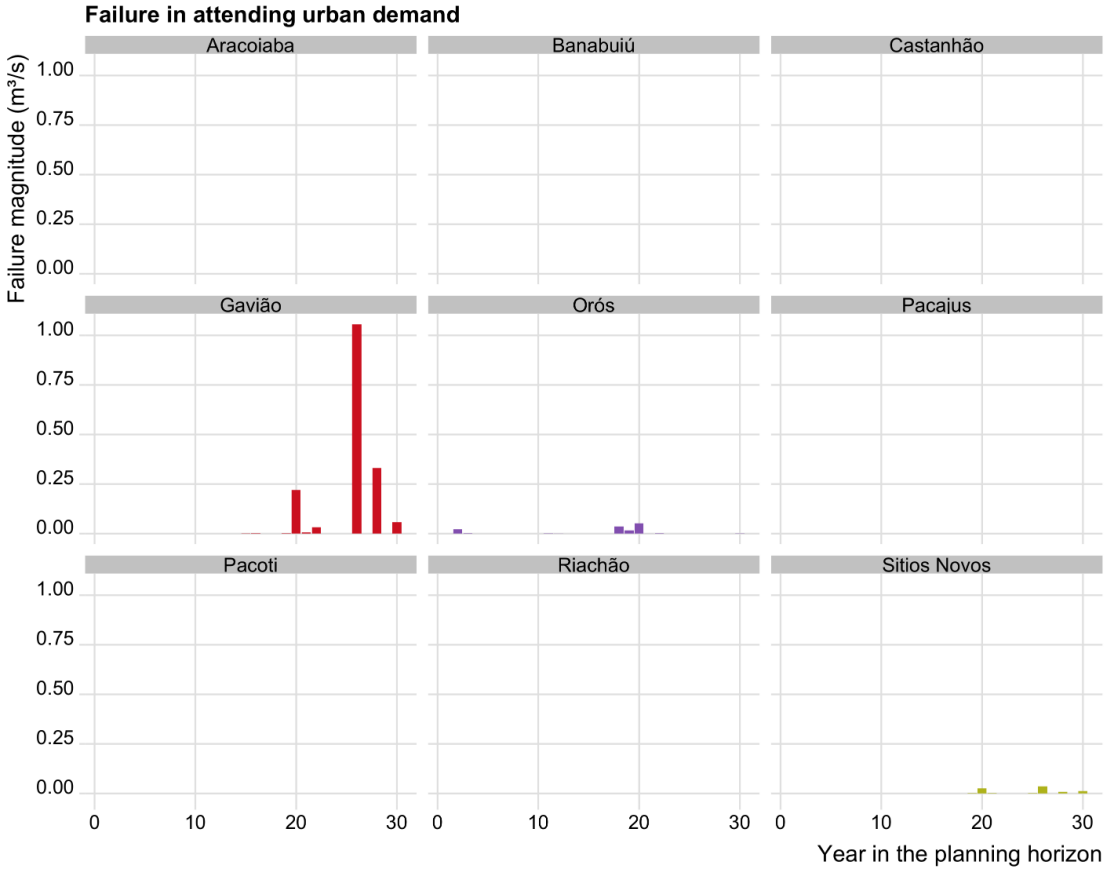
Figure 33 – Failure frequency in attending urban water demand (grey bars) over the years in the planning horizon. Colored lines indicate the ECDF of the activation year of desalination, water transfer (PISF and wastewater reuse), calculated for 100 simulations of the optimal expansion strategy.



Source: The author.

Water demand for irrigation, which was the second priority in relation to urban demand, was not completely met in the last ten years of the planning horizon (Figure 35). This

Figure 34 – Failure magnitude (m³/s) in attending urban water demand (grey bars) over the years in the planning horizon, calculated for 100 simulations of the optimal expansion strategy.



Source: The author.

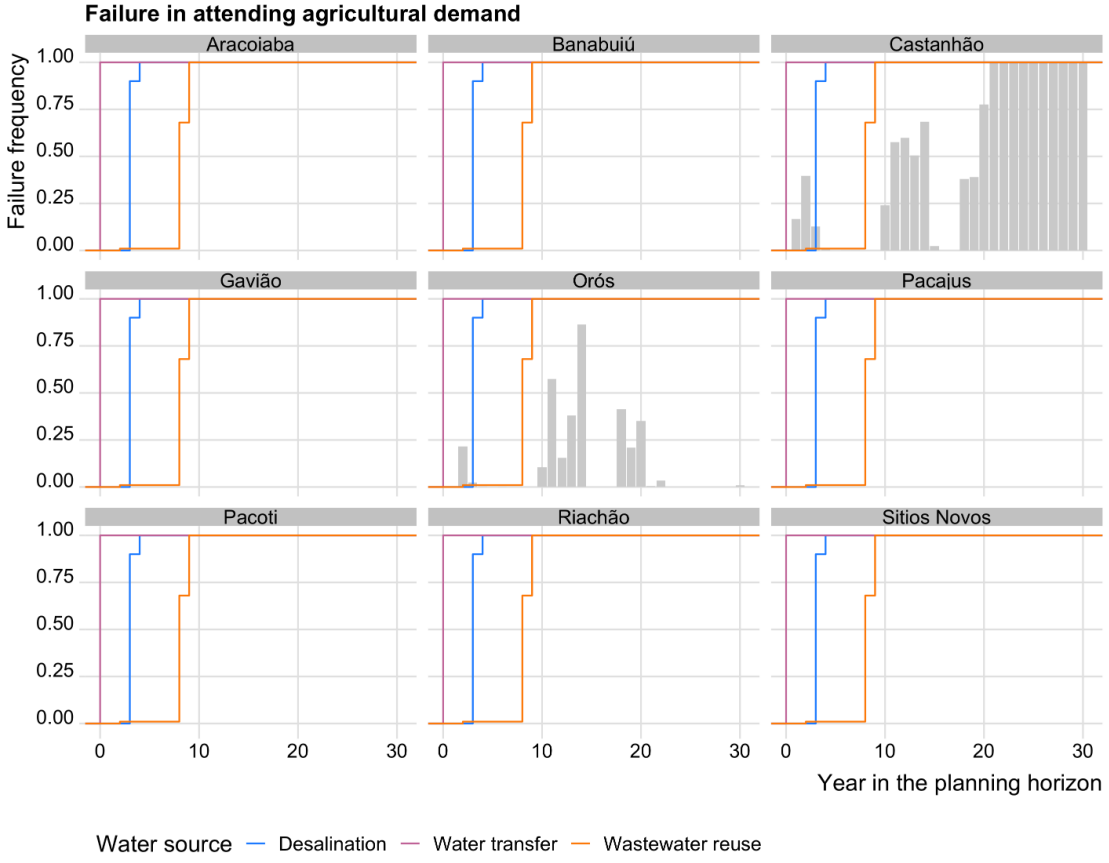
means that incorporating alternative water supply sources into the system might not be enough to ensure agriculture supply in the Jaguaribe region, but it can guarantee urban water supply in both Jaguaribe and Metropolitana water basins for at least 30 years.

Hence, decision makers and stakeholders should focus on adopting measures to reduce future water demand (both urban and agricultural), such as encouraging households to install water saving devices (e.g. dual flushers, low flow showers) (ABU-BAKAR *et al.*, 2021), replacing inefficient pumps (URRESTARAZU; BURT, 2012), encouraging organic farming (WHEELER *et al.*, 2015) and setting limits to water allocation (PERRY *et al.*, 2017). The benefits of increasing irrigation efficiency are extensively discussed in the literature (ADAMSON; LOCH, 2014; OECD, 2016), and although it does seem to drive an increased water consumption and only a modest increase in water productivity (WARD; PULIDO-VELAZQUEZ, 2008), it can be helpful if combined with a water allocation scheme based on an accurate water accounting system (PERRY *et al.*, 2017).

It is interesting to note that water transfer from PISF is not used at the end of

the planning horizon, even though water is needed to supply agricultural demand 37. One possible reason for that is the optimization algorithm itself, which minimizes immediate costs and expected future costs associated with the decision to be taken in the current state. This means that by the end of the planning horizon, it might not be worth it to fully attend agricultural demand in order to guarantee water availability in future stages. This is also a consequence of the penalties attributed to each water use type.

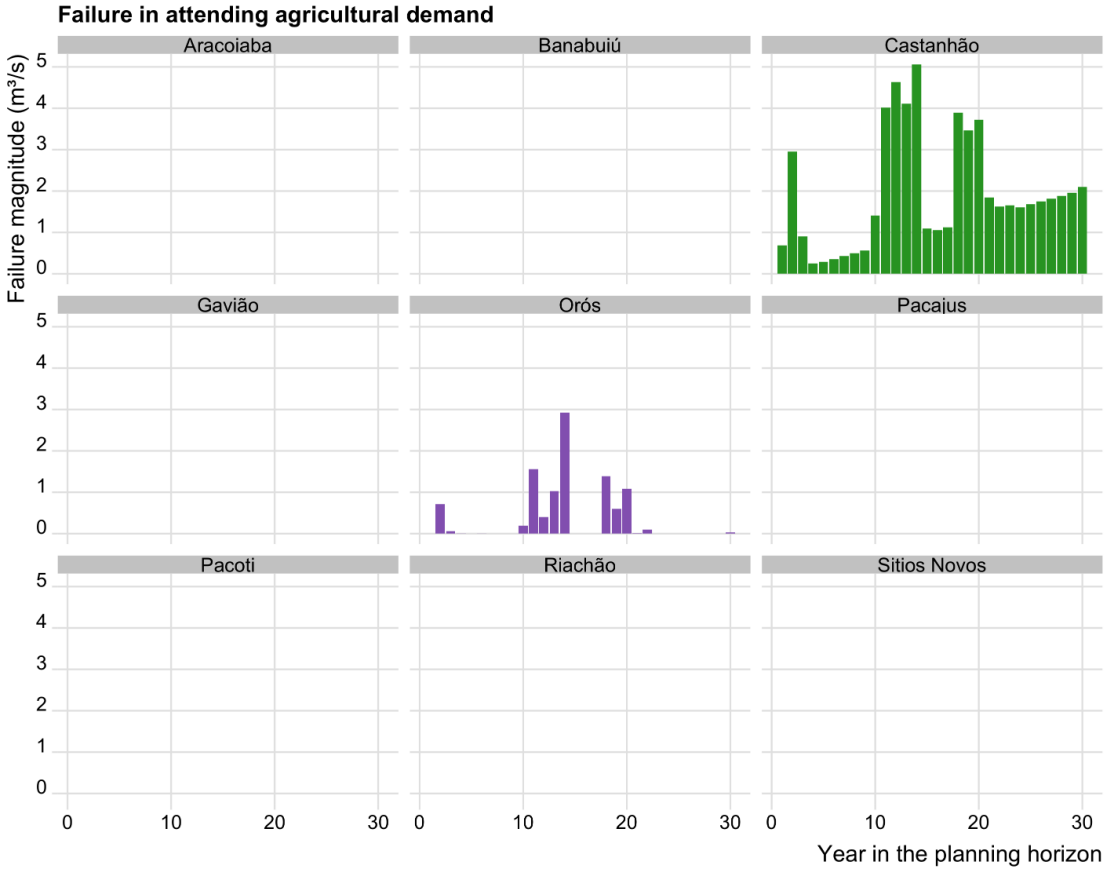
Figure 35 – Failure frequency in attending agricultural water demand (grey bars) over the years in the planning horizon. Colored lines indicate the ECDF of the activation year of desalination, water transfer (PISF and wastewater reuse), calculated for 100 simulations of the optimal expansion strategy. Only Castanhão, Orós and Banabuiú had agricultural water demands associated with them.



Source: The author.

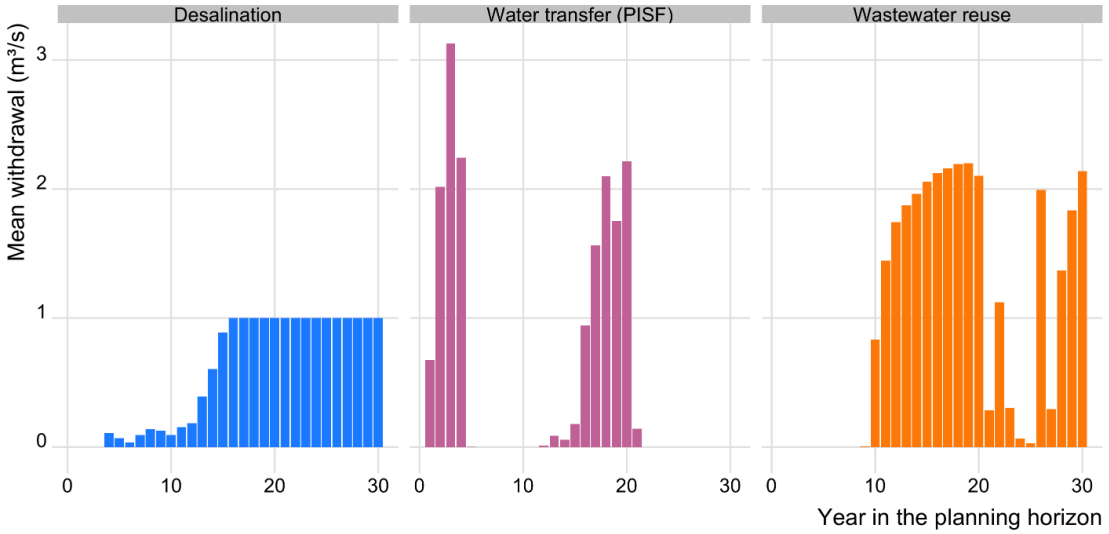
Some additional factors might affect optimization results, such as the initial reservoir volume (fluctuations above 10% of total storage capacity), penalty values (for not supplying water demands), the minimum reasonable storage volume, and the probability distribution of inflow. Analyzing the effects of these assumptions was beyond the scope of this study.

Figure 36 – Failure magnitude (m³/s) in attending agricultural water demand (grey bars) over the years in the planning horizon, calculated for 100 simulations of the optimal expansion strategy.



Source: The author.

Figure 37 – Withdrawals from the alternative water sources to be included in the supply system of the RMF.

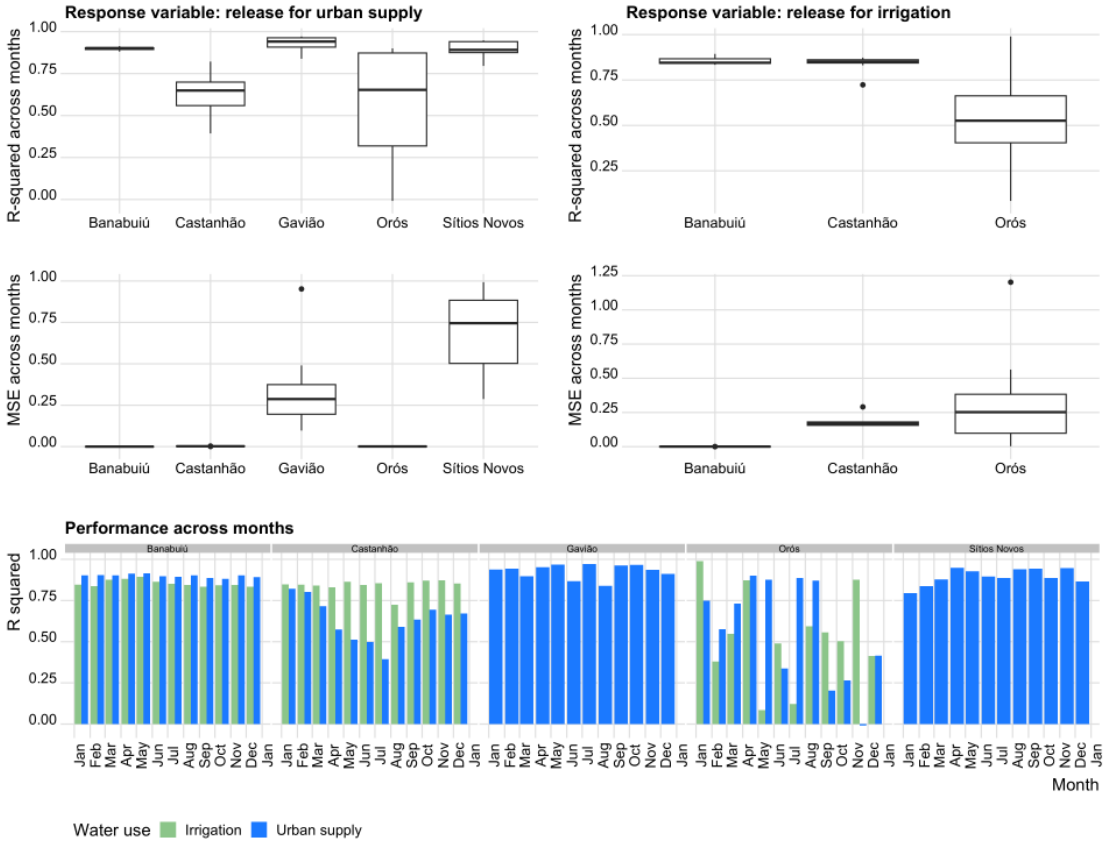


Source: The author.

7.3.2 Operating rules extraction

We obtained the reservoir operating rules in the following sequence: (i) first, we fitted the regression tree model for the Castanhão reservoir, using, besides its own storage, Orós storage (which is upstream) as predictors, and the release for urban supply as a response; (ii) then, we fitted a regression tree model for Orós reservoir, based on its own storage and the releases for irrigation and urban supply from Castanhão; (iii) finally, we obtained the regression tree model for all other reservoirs, using the storage information available from other reservoirs upstream. We repeated this procedure considering both the release for (i) urban supply and (ii) irrigation as response variables, but we considered the release for urban supply as an input variable for the irrigation regression trees. We only considered instances where both desalination and wastewater reuse were already being used to supply water.

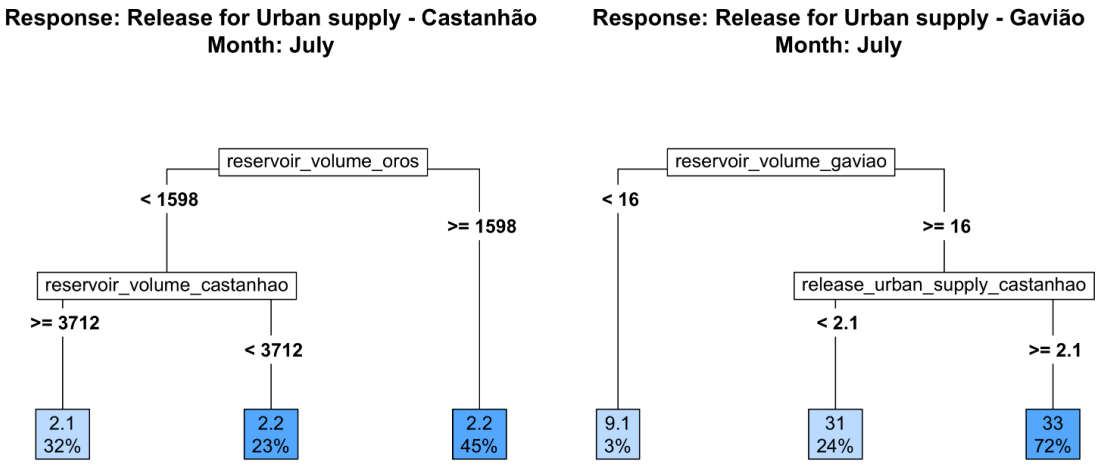
Figure 38 – Performance of regression tree models representing the operating rules of the reservoirs of the JMS. In the right panel, we present the performance metrics for the models where the response variable was the release for irrigation, and in the left, models where the response was the release for urban supply. Below, we present how the R^2 varied across months for both models.



Source: The author.

Then, we calculated R^2 and MSE for all obtained models to assess how representative they are of the optimal reservoir operating strategy (Figure 38). While the operation of Banabuiú, Gavião and Sítios Novos was well represented by the regression tree models, release for urban supply by Castanhão had less satisfying estimations from March to July (overlapping with the rainy season). Assessment of the release for both urban supply and irrigation in Orós has a fluctuating performance across months.

Figure 39 – Regression trees obtained to estimate the release for urban supply in July for Castanhão (left) and Gavião (right) reservoirs.



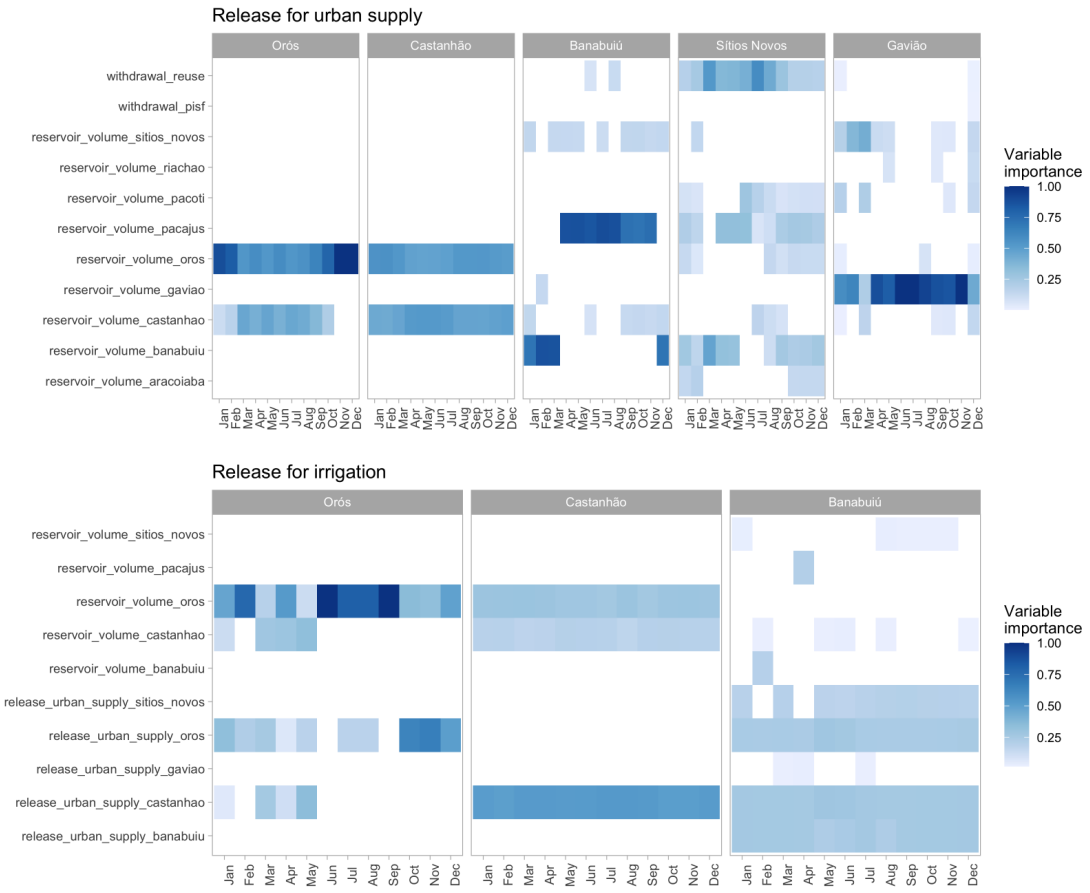
Source: The author.

Figure 39 presents two regression trees obtained for the Castanhão (Jaguaribe basin) and Gavião (Metropolitana basin) reservoir. The advantage of this model is that it is easy to interpret and can be helpful for stakeholders to make decisions about water allocation strategies (SOUZA FILHO *et al.*, 2023). All other operating rules obtained with the regression tree model are attached to this thesis (Appendix 11).

It is also insightful to note which predictors have the strongest impact on the release decisions. To do that, we calculated the variable importance measure for each model 40 and normalized it to compare their influence on the operation decisions. Variable importance is the decrease in the MSE is obtained by adding a split based on that variable to the tree. Overall, water stored in the reservoir at the beginning of the month is the most relevant variable for deciding how much water to release for urban supply. However, for the Sítios Novos reservoir, which supplies the main industrial demand of the RMF, the release depends heavily on how much water comes from wastewater reuse. Decisions regarding the release for irrigation depend

mainly on the release for urban supply, and indirectly, on reservoir storage.

Figure 40 – Importance of predictors for each regression tree model. On the top, it the reservoir for which the release prediction is made; on the y axis, are the predictor names. Variable importance was normalized for each month.



Source: The author.

7.4 Conclusion

Planning the expansion of a water supply system involves making decisions about how to manage new and current water sources over time under climate uncertainty. In this study, we approach this problem using a combination of optimization and machine learning techniques. First, we identify the best expansion strategy by formulating a multi-stage stochastic problem to minimize water supply failure and the costs of expansion. Then, we simulate the best policy and extract the reservoir operating rules from the results of several random streamflow realizations.

By applying this method in Fortaleza, Brazil, we find that installing both desalination and a wastewater reuse plant and transferring water from another basin will be essential to ensure water supply for the next 30 years. However, even by adding alternative sources to the

supply system, agricultural demand might not be fully attended by the end of the planning horizon. But the benefits of including alternative water sources into the supply system go beyond supplying water demands, and also include social and environmental gains. Wastewater reuse might have positive impacts such as undischarged pollution (GARCIA; PARGAMENT, 2015; HERNáNDEZ-SANCHO *et al.*, 2010) and indirect improvement of public health (BDOUR *et al.*, 2009). Although desalination might have increased environmental costs (e.g. disposal of highly concentrated brine, greenhouse gases emission due to increased energy consumption), its availability is not restricted by climate variability. Inter-basin water transfer can mitigate ecological water deficiency in the recipient basin and benefit fauna and flora, but on the downside, it can result in salinization of soil in the donor basin (ZHUANG, 2016). The ecological impacts of PISF must be further investigated.

One drawback of this study is that we do not assess the economic benefits of irrigation nor the environmental costs associated with desalination and wastewater reuse. We are mostly concerned with finding strategies to increase water supply under climate uncertainty in the long-term and analyzing the benefit of including alternative water sources in the water supply portfolio. Further research could address both environmental and social costs of including new water sources to the supply system and climate change effects on water availability.

To close the gap between water demand and supply and ensure water security under uncertain climate conditions, demand-side measures can be helpful. Price-based measures have been proved to be effective to reduce domestic water demand (), but also unfair to lower income residents (see Chapter 6). Customized demand-side programs (e.g. conservation feedbacks, implementing water efficient devices) designed for user profiles seem to be a better approach (COMINOLA *et al.*, 2018; QUESNEL; AJAMI, 2017).

The Master Plan for Agriculture in Ceará (FRUTAL, 2013) pinpoints that most irrigation systems in the state are technologically inadequate and use outdated irrigation methods. Improving such systems might be necessary to reduce agricultural water use, but practitioners should keep in mind that it may also increase water consumption (BERBEL *et al.*, 2018). To minimize this effect, water managers should devote to refine the water accounting system and couple it to the allocation and permit granting policies.

Losses in water distribution networks, specially physical and/or real can also account for a major portion of the imbalance between demand and supply. Total water losses in Ceará were around 44.9% in 2022 (OLIVEIRA *et al.*, 2022), hence water pressure management

(KARADIREK *et al.*, 2012) and efficient loss detection (ADEDEJI *et al.*, 2017) should be a priority for the water managers in the state.

This study provided a method to analyze long-term planning of a water supply system and a strategy to guide operational decisions by using an interpretable machine learning model. The proposed approach is useful for detecting potential expansion strategies and the implications of the growing water demands.

8 UNCOVERING THE INFLUENCE OF HYDROLOGICAL AND CLIMATE VARIABLES IN CHLOROPHYLL-A CONCENTRATION IN TROPICAL RESERVOIRS WITH MACHINE LEARNING

> Viver - não é? - é muito perigoso. Porque ainda não se sabe. Porque aprender a viver é que é o viver, mesmo. O sertão me produz, depois me engoliu, depois me cuspiu do quente da boca... O senhor crê minha narração? (ROSA, 2019)

8.1 Introduction

In most developing countries, the urbanization process is associated with an increase in water demand (UNESCO, 2018). At the same time, the availability of drinking water remains the same or even decreases (VELDKAMP *et al.*, 2017; GREVE *et al.*, 2018). Accelerated urbanization is also related to the intensification of human activity, resulting in increased nutrient loads and water quality degradation (VöRöSMARTY *et al.*, 2010).

The situation is worse in regions with high climatic variability (temporal and spatial), in which the distribution of rainfall is irregular, and extreme events of droughts and floods are frequent (EASTERLING *et al.*, 2000; HIRSCH; ARCHFIELD, 2015). This is the case in the Northeastern semi-arid region of Brazil, where multi-annual drought events are common and have severe socioeconomic and environmental impacts (CAMPOS, 2015; PONTES FILHO *et al.*, 2020). One of the management strategies historically adopted in the region to deal with this scenario is the construction of reservoirs (GUTIéRREZ *et al.*, 2014), which have the important role of transferring water both temporally and spatially. Most of these reservoirs serve multiple purposes, including drinking water supply, irrigation, and fish farming. The water volume in these reservoirs can vary significantly between the dry and wet seasons and reduce drastically during drought periods (ROCHA; LIMA NETO, 2021b).

Eutrophication, caused by the excessive increase of phosphorus and nitrogen loads, is one of the main causes of the deterioration of water quality in reservoirs (PAERL; OTTEN, 2013). Eutrophication is associated with the proliferation of algae and cyanobacterial blooming (YANG *et al.*, 2008), and sometimes, an increase in mortality of benthic animals and fish (SPERLING, 2005). Agriculture and livestock farming contribute to this process since significant loads of phosphorus and nitrogen can be carried with surface water runoff into the reservoir (WIEGAND *et al.*, 2020; ROCHA *et al.*, 2020; ROCHA; LIMA NETO, 2022a).

A few studies have associated phytoplankton growth rates with the volume of water stored in the reservoir (PACHECO; LIMA NETO, 2017; JUNIOR *et al.*, 2018a), but most of them relied on field studies, which are usually unavailable for a long-term horizon (more than 10 years), especially in data-scarce regions. Other researchers have related Chla to hydrological and/or climate variables, such as wind speed, air temperature, solar radiance, precipitation, mixing depth, and runoff (BLAUW *et al.*, 2018; STOCKWELL *et al.*, 2020; STEFANIDIS *et al.*, 2021), but none of them analyzed this relationship in tropical reservoirs. Past research has also shown that climate variability and future changes in frequency and intensity of drought events can increase phosphorus concentrations in tropical reservoirs (RAULINO *et al.*, 2021; ROCHA; LIMA NETO, 2021b), hence the importance of investigating the relationship between climate variables and Chla.

The mechanisms associated with Chla fluctuations are complex and have been extensively studied (PACHECO; LIMA NETO, 2017; BLAUW *et al.*, 2018; DUNSTAN *et al.*, 2018; LI *et al.*, 2021), and more recently, many researchers have applied machine learning techniques for water quality assessment and to predict Chla (LIU *et al.*, 2019; SHEN *et al.*, 2019; AHMED *et al.*, 2019; TONG *et al.*, 2019; MAMUN *et al.*, 2019; NGUYEN *et al.*, 2020; YU *et al.*, 2020). Data for most of these studies have been obtained from automated stations (BLAUW *et al.*, 2018) or long field campaigns (LIU *et al.*, 2019; AHMED *et al.*, 2019; LI *et al.*, 2021), which can be expensive and time consuming. One strategy to deal with the lack of field data is using satellite data, which has been frequently used to monitor water quality and has proved to be reliable, but it has not been sufficiently explored for inland waters (LOPES *et al.*, 2014; GHOLIZADEH *et al.*, 2016; WANG; YANG, 2019; ROSS *et al.*, 2019; NGUYEN *et al.*, 2020; IIAMES *et al.*, 2021).

Recent evidence suggests that reanalysis climate data can be effective in explaining the effects of climate on phytoplankton biomass (STEFANIDIS *et al.*, 2021). However, to the authors' knowledge, no study has explored the predictive capacity of non-parametric models based on reanalysis climate data for semiarid climates. In these regions, Chla modeling can be challenging, as water volume has a strong interannual variability and phosphorus concentration has a weak correlation with Chla. The state-of-the art models used to explore the mechanisms for Chla variability may not be suitable for them. Machine learning models can be informative in this case, but model comparison is required, as these algorithms are mainly driven by data and their predictive capacity can be site-specific.

This study evaluates the influence of hydrological and climate variables on Chla in reservoirs located in Northeastern semi-arid Brazil. This analysis is important from the point of view of climate variability, which can significantly affect the hydrological processes of the reservoirs, and to understand the possible influence of water level and volume fluctuations on Chla. The predictive model proposed here combines climate reanalysis data, together with commonly available hydrological variables, and satellite-based predictions of Chla. The main goals of this study are (i) to explore the relationships between hydrological and climate variables and the concentration of Chla in tropical reservoirs, and (ii) to evaluate the performance of nonparametric machine learning models for predicting Chla using these variables.

8.2 Methodology

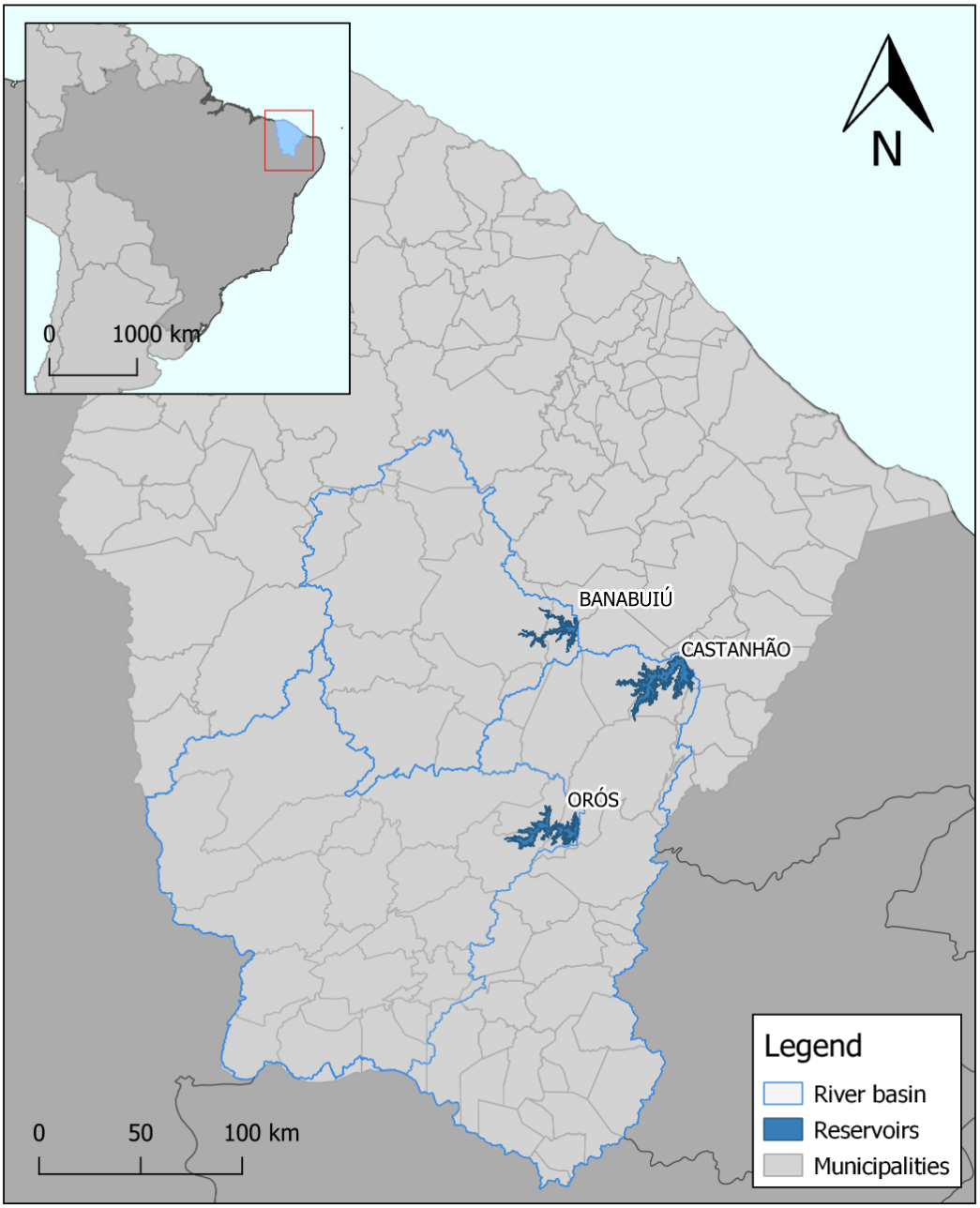
The reservoirs analyzed in this study are located in the Northeastern region of Brazil (Figure 41), which has a semi-arid climate and is frequently affected by multi-annual droughts. These reservoirs are part of the JMS, which transfers water to Fortaleza, the capital of the State of Ceará. Castanhão is the largest reservoir for multiple uses in the country, with a capacity of 6.7 billion cubic meters. All three reservoirs are also used for irrigation. Banabuiú (capacity of 1.6 billion cubic meters) supplies the Irrigated Perimeter Morada Nova, while Orós (capacity of 2.1 billion cubic meters), the second-largest reservoir in the State of Ceará, also serves for hydroelectric use. The surface area of these reservoirs ranges between 116 and 410 km^2, and the mean water level from 90 to 192 m.

8.2.1 Data and variable selection

This research uses data from publicly available databases, obtained from satellite, reanalysis, and rain gauge stations. The historical series of monthly Chlorophyll-a concentrations (Chla) from 2002 to 2019 were obtained from the Hidrosat portal (http://hidrosat.ana.gov.br/). The dataset obtained from Hidrosat is the result of a partnership between the Brazilian Water and Sanitation Agency (ANA) and the *Research Institute for Development* / Institut de Recherche pour le Développement (IRD). Water quality stations use data from the Terra (EOS AM) and Aqua (EOS PM) satellites.

The program MOD3R (MODIS Reflectance Retrieval over Rivers) is used to extract time series of reflectance from MODIS (sensor onboard the Terra and Aqua satellites) images of

Figure 41 – Study area location. Banabuiú, Castanhão, and Orós are the main reservoirs of the State of Ceará, Brazil (highlighted in the map). Their hydrographic basins are contoured by the blue line.



Source: The author.

water bodies. The algorithm identifies and groups the water pixels in the image and, from the extraction of reflectance values from the visible and infrared bands, the water quality parameters are estimated. Mathematical models that relate reflectance data and water quality data were calibrated and validated with data collected in the field. This procedure is detailed in Lins *et al.* (2017).

For some months of the original series of Chla, more than one estimation was

available. In these cases, the median of these values was used to represent monthly concentration. Months with missing values were filled in with the median of the historical concentration series for the corresponding month. Hydrological and climate variables used in this research and their respective sources are described in Table 23. Precipitation data for the period between 2002 and 2019 were obtained from the spatial interpolation of the data provided by the Brazilian Water Agency, publicly available on the Hidroweb portal (http://www.snirh.gov.br/hidroweb/). Daily precipitation measured in rain gauges was interpolated using the inverse distance weighting method with exponent two into grid points with 0.05° size. This procedure was performed using the R package ipdw (Stachelek 2020). Then, the average monthly precipitation was calculated for each reservoir's hydrographic basin.

Average monthly temperature data was extracted from version 4 of the University of East Anglia's CRU climate database (HARRIS *et al.*, 2020). Data is publicly available in the NetCDF format, which stores multidimensional variables; for example, temperature has four dimensions: latitude, longitude, time, and temperature value. To estimate average monthly temperature over the reservoir, we extracted the pixels contained inside the limits of the reservoir and calculated its average value for each month in the time series (2002-2019).

Except for water volume and level, all other variables were extracted from the ERA5 gridded (lat-lon grid of 0.25 degrees) reanalysis database of the European Center for Medium-Range Weather Forecasts (HERSBACH *et al.*, 2020). Data is also available online in the NetCDF format, in hourly or monthly scale, with a temporal coverage from 1979 to present. Reanalysis uses observed data from weather stations across the world and climate models to estimate a global dataset containing atmospheric, land and oceanic climate variables.

Average runoff was calculated by averaging the monthly runoff for all pixels contained in the region delimited by each reservoir's hydrographic basin. For all other variables, the time series was extracted for the nearest pixel to the centroid of the reservoir, which was identified using the nearest-neighbor interpolation method. Water volume and level were obtained from the COGERH, also available online on the Reservoir Monitoring System (https://www.ana.gov.br/sar).

Further improvements can be made by validating reanalysis data with field data and by incorporating more reservoirs into the analysis. However, this would require field campaigns and/or the implementation of automatic monitoring systems.

Variables that had a Pearson correlation coefficient above 0.8 were removed from

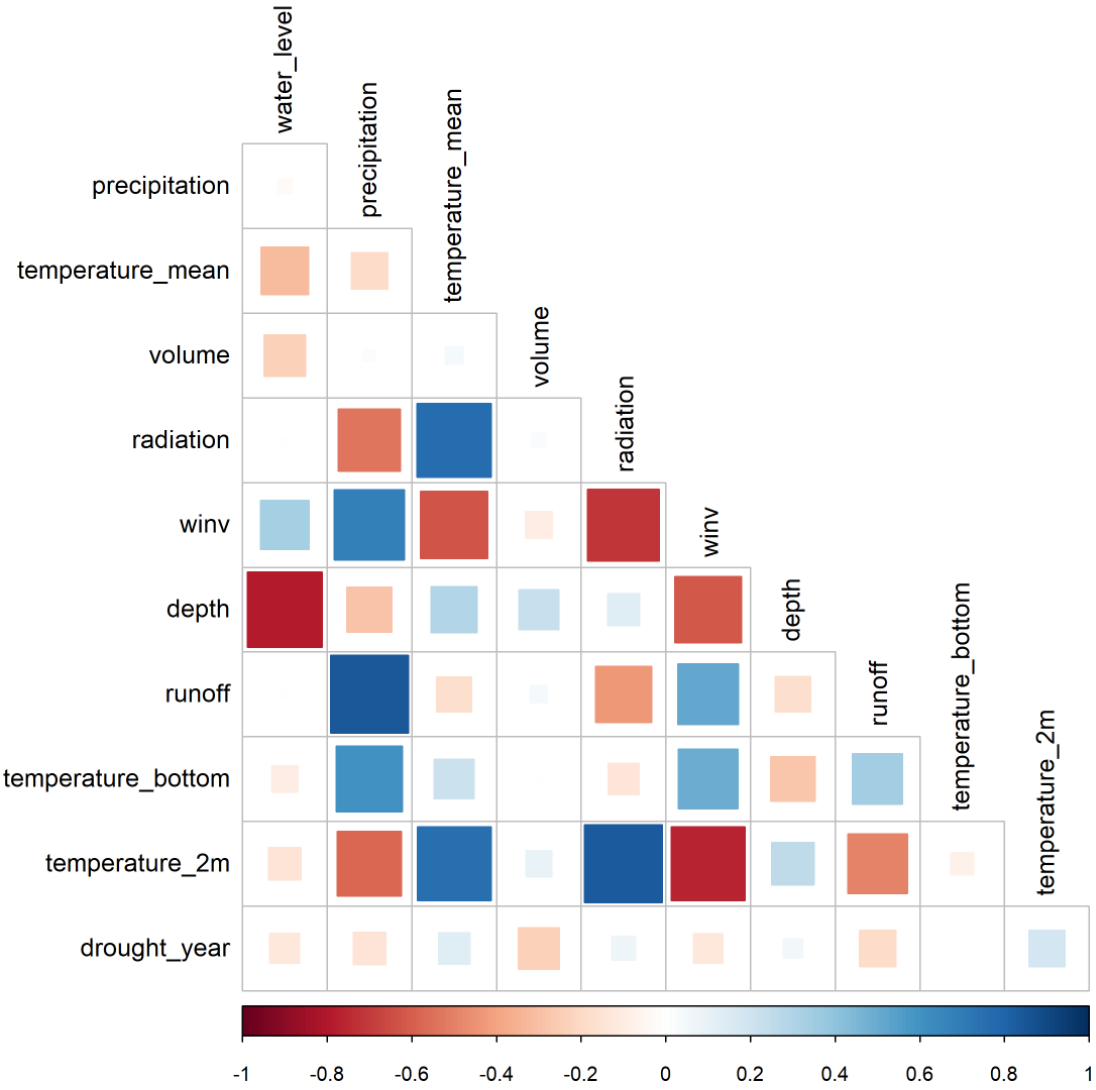
Table 23 – Explanatory variables of the regression models.

Variable	Unit	Description	Source	Mean	Standard Deviation
Mean precipitation	mm	Average monthly precipitation on the hydrographic basin of the reservoir, calculated from rain gauge measures	Hidroweb*	61.59	74.74
Mean temperature	C	Mean surface temperature over the reservoir calculated from CRU grid	CRU (HARRIS *et al.*, 2020)	27.78	1.23
Water volume	m^3	Total water volume in the reservoir	COGERH*	1.42E+09	1.56E+09
Water level	m	Distance from the bottom of the reservoir to the water surface	COGERH*	137.18	43.99
Runoff	m	Monthly average of surface and subsurface runoff accumulated over one day in the hydrographic basin	ERA5	1.75E-04	3.59E-04
2m temperature	K	Air temperature at 2 m above the reservoir	ERA5	300.90	1.29
Lake bottom temperature	K	Water temperature at the bottom of the reservoir	ERA5	299.00	1.20
Lake mix-layer depth	m	Thickness of the uppermost layer of the reservoir that is well mixed and has a near constant temperature	ERA5	5.19	1.48
Surface net solar radiation	Jm^{-2}	Amount of solar radiation that reaches the water surface, assuming cloudless conditions	ERA5	1.78E+07	2.42E+06
10m u-component of wind	$ms-1$	Horizontal wind speed of air moving towards the east, at a height of 10 m above the reservoir surface	ERA5	-2.43	0.80
Reservoir	Dummy	Represents the reservoir correspondent to the observation	-	-	-
Drought year	Binary	Indicates if the year of the observation was a drought year (1) or not (0)	-	0.37	0.48

Source: The author.
Note: Except for the variables extracted from the sources indicated with an asterisk "*" (which are available in tabular format), all other variables were obtained in NetCDF format.

the dataset (temperature at 2 m and runoff; refer to Figure 42 in the supplementary material for the correlation matrix). As the effect of hydrological variables can be site-specific, a dummy variable was included to indicate the corresponding reservoir of each observation. To account for the effect of drought on Chla, a binary variable was included to indicate if the observation was registered during a drought year, according to drought records of the area (PONTES FILHO *et al.*, 2020).

Figure 42 – Pearson correlation coefficient between explanatory variables.



Source: The author.

All explanatory variables were re-scaled to range between 0 and 1 using the min-max normalization:

$$x' = \frac{x - min(x)}{x - min(x)}$$

Where x is the original value and x' is the scaled value. The final dataset contained 679 samples from the three reservoirs analyzed in this study. All analyses were performed using R (version 4.0.5) software.

8.2.2 Regression models

Six nonparametric machine learning models were compared with standard linear regression and one semi-parametric algorithm to investigate the best-performing predictive model. Data were randomly split into training (80%) and testing (20%) datasets. The training dataset was used to tune model hyperparameters, and the testing dataset was used to evaluate model performance. Model tuning and performance evaluation are detailed in section 2.4.

In the following topics, there is a brief explanation of the regression models used in this study. It is important to highlight an essential property of the predictive models, which is the bias-variance tradeoff. When fitting regression models, the best outcome is obtaining a model that not only provides accurate predictions (low bias) but also generalizes well to new data (low variance). The bias error is associated with a poor learning process, in which the relationship between explanatory and response variables is not properly captured (underfitting). The variance error happens when the model is sensitive to small variations during training, i.e., fits too perfectly and ends up modeling random noise (overfitting). One wants to avoid models that are either too complex or too simple and get the one that presents similar performances during training and testing.

8.2.3 Linear Regression Model

Linear regression aims to explain the relationship between a set of independent variable vectors (x) and a dependent variable (y) based on the linear function described below:

$$\hat{Y} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

Where X_j is a vector for the jth independent variable, and β_j and β_0 are unknown parameters (coefficients and an intercept, respectively). The algorithm calculates the parameters by minimizing the sum of the squares of the residuals (SSR), i.e., the difference between observed and predicted values.

8.2.4 Elastic-Net Regularized Generalized Linear Model

While in the ordinary least squares regression the distribution of errors is normal, in the GLM, it may assume different distributions, such as Binomial, Poisson, and gamma. In GLMs, the variance of the response variable can be non-constant and a linking function

can be used to connect the predictor and the mean of the distribution function (NELDER; WEDDERBURN, 1972). In this study the error distribution was assumed to be normal.

Regularization is a useful technique for learning algorithms: penalties can be added to the model to prevent overfitting issues and to deal with highly correlated explanatory variables. Ridge and Lasso regression are some of the simplest and widely used penalized models; they work by adding a penalty to the SSR. Lasso penalizes the sum of the absolute coefficients (ℓ_1 penalty) and might lead to variable selection as it sets coefficients to zero if λ is sufficiently large. The parameter λ controls the regularization strength and might assume any positive value.

$$SSR_{lasso} = \sum_{j=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Where y_i is the observed value, $\hat{y}i$ is the predicted value, n is the number of samples, β is the coefficient vector, and p is the number of explanatory variables. Ridge regression penalizes the square of the magnitude of the coefficients (ℓ_2 penalty) and shrinks the coefficients proportionally, keeping all of the variables in the model:

$$SSR_{lasso} = \sum_{j=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

The linear combination of both penalties is called elastic net regularization, controlled by the parameter α, which ranges between 0 (ridge) and 1 (lasso).

8.2.5 Artificial Neural Network

An artificial neural network is composed of interconnected nodes (or neurons) arranged in layers (HASTIE *et al.*, 2009). The MLP, a broadly used class of neural networks, consists of the input (which receives the independent vectors), output, and one or more hidden layers. These layers have weighted connections that are adjusted as training occurs and are fully connected, i.e., a neuron in one layer is connected to every neuron in the next layer. The number of neurons in the hidden layer is critical for the learning process, as they detect the characteristics present in the training data and apply a nonlinear transformation to the input data.

The training algorithm used in this study was the backpropagation of the error, in which the gradient of the error concerning the weights is calculated layer by layer. Then, the error is calculated, and all weights are updated backward through the network. The optimization algorithm used to perform this method was gradient descent.

An MLP with a single hidden layer was selected and the number of hidden nodes was adjusted in the training process (see Table 24). The number of nodes in the input layer was set to 10 (the number of explanatory variables), and the learning rate was set to 0.1.

8.2.6 k-Nearest Neighbors

The k-Nearest Neighbors (kNN) is a supervised algorithm (ALTMAN, 1992) for classification and regression based on a similarity measure, such as distance functions. In this method, one finds the k observations in the training set closest to x and (i) average their responses, for regression tasks or (ii) take the majority class among its k nearest neighbors, for classification tasks. The equation for the kNN fit for Y can be described as:

$$\hat{Y}_x = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

Where N_k is the neighborhood of x defined by the k closest points x_i in the training sample. The only parameter to be determined is the number of neighbors k.

8.2.7 Classification and Regression Tree

A decision tree provides a set of rules to express the relationship between explanatory and response variables, which are represented with a tree structure. The leaves represent class labels (classification), or estimations of the response variable (regression), and branches represent the values of the tested variable.

Regression trees predict using the average values of y within each subset, which is selected to minimize the MSE. To determine whether splitting should continue to be done, one can use some combination of (i) a minimum number of points in a node, (ii) purity or error threshold of a node, or (iii) maximum depth of the tree (KRZYWINSKI; ALTMAN, 2017). Here, the minimum number of points per node was set to 20. The complexity parameter, which corresponds to the minimum improvement in the model needed at each node, was tuned using grid search (see Table 24).

8.2.8 Tree-based Ensemble Models: Random Forest and Gradient Boosting Regression

Decision trees alone can easily overfit, depending on the size of the training dataset. An ensemble of decision trees is an effective approach to build a robust model and prevent overfitting. RF combine shallow trees using bagging, i.e. the prediction is the average (for

regression) or the majority vote (classification) of the trees in the ensemble (BREIMAN, 2001). The trees are constructed from bootstrap samples and a random subset of predictors (mtry) is used at each split in a tree. Together with the number of trees, these are the main parameters of random forests, which was tuned in the training process (see Table 24). The minimum number of observations per node was set to 20. GBM uses a different ensemble technique called boosting, where decision trees are combined in a forward stage-wise procedure. While in RF each tree is independently built, in gradient boosting, each new tree is constructed on the residuals of the previous tree to minimize the Mean Squared Error. The maximum depth of the trees (interaction depth) was tuned between 1 to 6, while the minimum number of observations per node was set to 10. The values set for the other parameters of GBM are described in Table 24.

8.2.9 Support Vector Machine

(BOSER *et al.*, 1992), although widely used for classification problems, might also be applied for regression (SVR). In , the main goal is to find a hyperplane that fits the training data by minimizing the Euclidean norm of the coefficient vector. This model uses a kernel function to map input data to higher-dimensional spaces, where it can be linearly separable. In regression problems, a symmetrical "margin" is added around the estimated function, where the absolute errors should be equal or less than the maximum error ε (AWAD; KHANNA, 2015). is an optimization problem where the objective function minimizes the Euclidean norm of the function coefficients (w), while avoiding outliers:

$$min\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}|\xi i|$$

Subject to:

$$|y_i - w_i x_i| \leq |\xi i|$$

Where C is the cost parameter, which gives more weight to the function flatness and ξ is the slack variable and corresponds to the tolerable distance of outliers from the margin. A Radial Basis Function kernel was applied here, defined as:

$$K_{RBF}(x,x') = e^{-\gamma\|x-x'\|^2}$$

Where x and x' are samples in the input data and γ is a parameter related to the variance of the function. This parameter was set to the inverse of the training data size.

8.3 Model parameters and performance evaluation

The tuning process of the hyperparameters of regression models is fundamental to avoiding overfitting. One of the most traditional approaches to optimize hyperparameter selection is grid search. In grid search, the modeler defines a subset of hyperparameter values and a performance metric to search for the best combination of parameters. Then, k-fold cross-validation or leave-one-out cross-validation can be used on the training set to perform the tuning process.

In this study, the RMSE was chosen to tune the model's parameters. Tuning was performed with a 5-fold cross-validation. In this approach, the training dataset is split into five subsets: the predictive model is fitted for four of them and the performance metric (in this study, RMSE) is calculated for the remaining subset. This procedure is repeated five times, so that all data is used at least once to train/validate the model. Model performance is assessed by calculating the average RMSE obtained in each subset. 5-fold cross-validation was applied using the R package 'caret'. Table 24 summarizes the main parameters of the fitted models and their correspondent values. Validation was performed for each combination of the parameters and the model with the best performance (lower RMSE) was selected.

8.4 Performance metrics

Model performance in the testing dataset was evaluated using the RMSE, MAE and the R^2 measures:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|^2}{n}}$$

$$MAE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n}$$

$$R_2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}$$

where y is the observed Chla, y is the predicted Chla, y is the mean observed Chla and n is the number of observations in the testing dataset.

Table 24 – Main parameters of the regression models used in this study. The values used to tune the models are indicated, and the chosen values are highlighted in bold.

Model	Main parameters	Values
Linear Regression Model	Intercept	**True** or False
Regularized Generalized Linear Model	Alpha	**0.10**, 0.28, 0.46, 0.64, 0.82, and 1.00
	Lambda	0.0046, 0.0173, 0.0646, 0.2409, **0.8979**, and 3.3469
Multilayer Perceptron	Number of nodes in the hidden layer	**3**, 5, 10 and 20
	Decay	**0.5**, 0.1, 1e-2, 1e-3, 1e-4, 1e-5, 1e-6 and 1e-7
k-Nearest Neighbors	Number of neighbors (k)	5, 7, **9**, 11, 13, and 15
Regression Tree	Complexity Parameter (cp)	**0.0274**, 0.0342, 0.0390, 0.0773, 0.1400, and 0.2066
Random Forest	Number of variables for splitting (mtry)	2, 4, 6, 8, 10, and 12
	Number of trees	50, 100, 250, 300
Gradient Boosting	Shrinkage	0.1
	Interaction depth	1, 2, 3, 4, 5, and **6**
	Minimum observations in node	10
	Number of trees	**50**, 100, 150, 200, 250, and 300
Support Vector Machine	Cost parameter (C)	0.25, 0.50, 1.00, 2.00, **4.00**, and 8.00
	Sigma	0.0619
	Epsilon (ε)	0.1

Source: The author.

8.5 Partial Dependence Plots

PDP were introduced by Friedman (2001) to interpret complex Machine Learning algorithms. The PDP represents the marginal effect of independent variables on the response of a machine learning model (FRIEDMAN, 2001). The partial dependence of the response on a variable x_l is represented by:

$$\hat{f}_{x_l}(x_l) = E_{x_s}[\hat{f}(x_l, x_s)] = \hat{f}(x_l, x_s)P(x_s)dx_s$$

Where x_l is the independent variable analyzed in the partial dependence plot, x_s is the subset of the other input variables of the regression model f and $P(x_s)$ is the marginal probability density of xs. The function shows the effect of the variable x_l on the dependent variable by marginalizing over the other explanatory variables.

8.6 Results

This section presents and compares the performance obtained with the predictive models, the relative importance of the hydrological and climate variables, and their relationships with Chla.

8.6.1 Performance of the regression models

Figure 43 presents the scatterplots of predicted and observed values for all the models tested in this study. From the plots, one can notice that linear regression, regularized GLM, and the regression tree underestimate Chla. These models have strong assumptions about error distribution: homoscedasticity, normal distribution, and no autocorrelation. Although the variables with an elevated correlation have been removed, there was still some multi-collinearity between the predictors, which could be a problem for the prediction. Predictors of water quality indicators will frequently be correlated (both temporally and spatially) since the mechanisms associated with their increase or decrease are interrelated (SU *et al.*, 2012; LIU *et al.*, 2019; MESQUITA *et al.*, 2020). It is important to keep in mind that highly correlated variables can present complementary information when combined (GUYON; ELISSEEFF, 2003), which reinforces the need for integrating correlation analysis with model-based variable importance.

RF, GBM, and MLP provided the best predictions (Table 25). These models are designed to capture nonlinear relationships between variables, which is likely to be the case here. RF and GBM can reduce the variance of the predicted values by employing ensemble techniques (boosting and bagging, respectively), outperforming the regression tree (HASTIE *et al.*, 2009). The model with a radial kernel is also able to detect nonlinearity, as it transforms data to a dimensional space where they can be linearly separable (AWAD; KHANNA, 2015). However, had a slightly worse performance than RF, GBM, and MLP.

As expected, the predictive models were able to explain only part of Chla, since the best performing model had an R^2 of 0.52 (Table 25). This performance can be considered satisfactory for a watershed-scale model, as a reference value to evaluate phosphorus (P) prediction (which can be easier to predict than Chla) is an $R^2 > 0.5$ (MORIASI *et al.*, 2015).

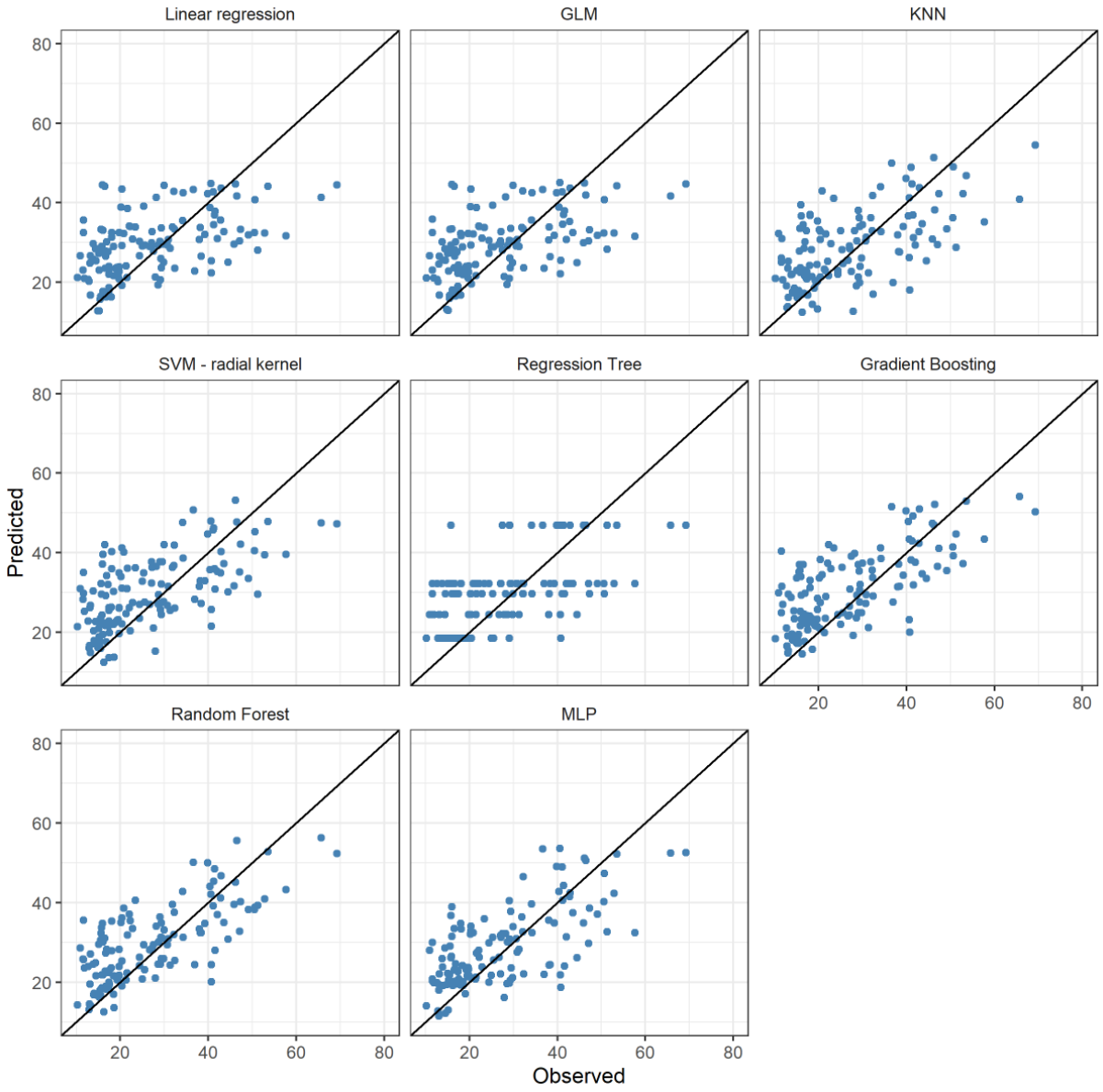
This result also suggests that hydrological and climate factors alone are not enough to predict Chla and additional variables might be necessary, such as water quality indicators (ROCHA *et al.*, 2020). However, it must be emphasized that the relationship between *P*

Table 25 – Performance metrics for the fitted models.

Model	R^2	RMSE	MAE
RF	0.52	9.32	7.15
GBM	0.46	10.26	8.01
MLP	0.45	9.74	7.66
	0.36	10.92	8.77
kNN	0.35	10.67	8.22
Regression Tree	0.32	10.77	8.21
Linear Regression	0.26	11.48	9.10
Regularized GLM	0.26	11.48	9.08

Source: The author.

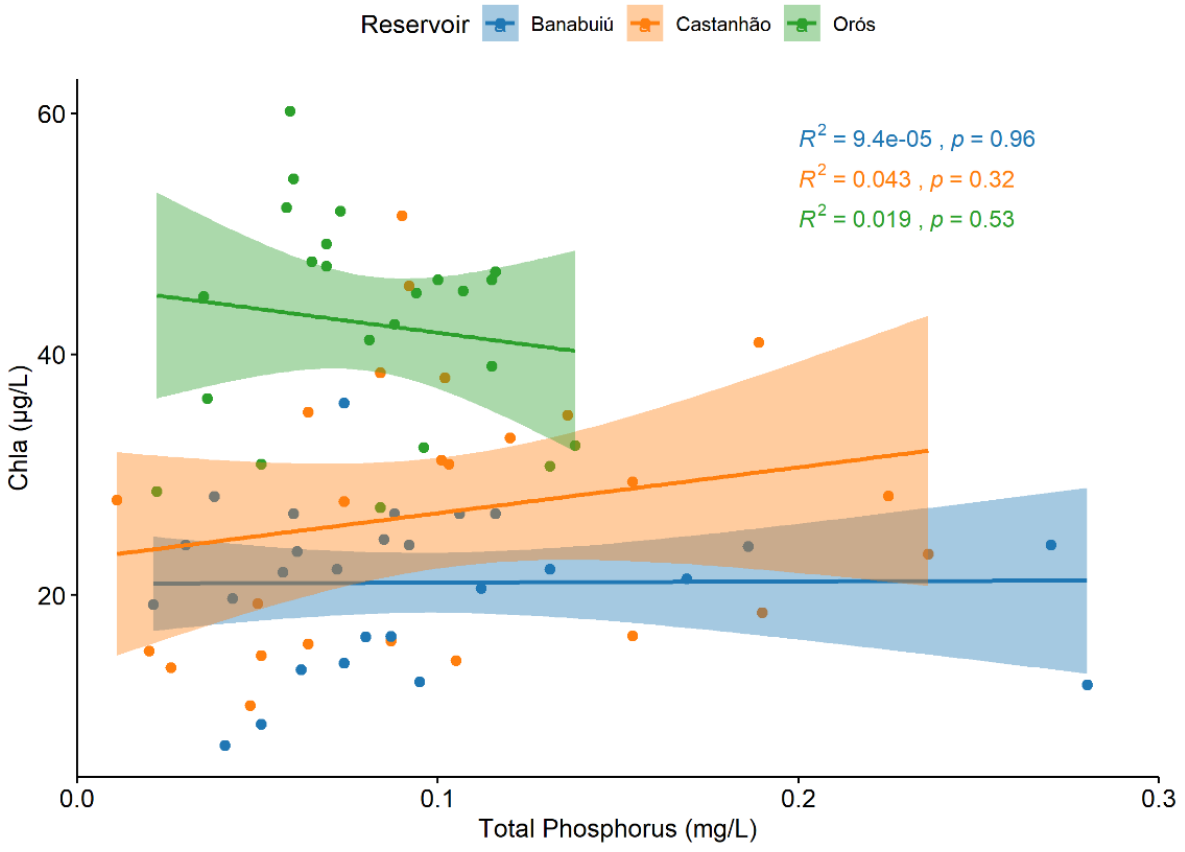
Figure 43 – Scatterplots for the predictive models tested in this study. The diagonal line represents the perfect fit between observed and predicted values.



Source: The author.

and Chla in tropical lakes is not comparable to that in temperate ones, where empirically estimated relationships between P and Chla provide reliable models to calculate Chla levels (SAKAMOTO, 1966; DILLON; RIGLER, 1974; JONES; BACHMANN, 1976). A correlation analysis between measured total phosphorus concentration, obtained from COGERH database (http://www.hidro.ce.gov.br/), and estimated Chla reveals that nutrient enrichment may not be the only influencing factor on eutrophication in tropical reservoirs (Figure 44). Although correlation between nitrogen and Chla was not analyzed here (since limited data was available), this can also be a limiting nutrient for eutrophication in reservoirs (QIN *et al.*, 2020).

Figure 44 – Correlation between total phosphorus and Chla in the reservoirs analyzed in our study. The dark, bold line represents the fitted regression line, and the shadow area is the confidence interval. Phosphorus measurements are taken each three months and were available for a shorter period than Chla estimations (05/2008 to 11/2019).



Source: The author.

Although past studies have obtained better predictive performances (STEFANIDIS *et al.*, 2021), Chla can be harder to predict in the semiarid, due to the significant water level variability (which implies more complex mechanisms behind eutrophication) and the usually higher trophic levels (WIEGAND *et al.*, 2021). There are, however, other possible explanations.

The Chla time series were derived from satellite data, which has high estimation accuracy (LINS *et al.*, 2017), but might contain noise or components that cannot be explained with known variables. Also, past studies have indicated that the drivers of Chla can vary with the temporal resolution (BLAUW *et al.*, 2018; LIU *et al.*, 2019). For example, on a monthly scale, water temperature is less important to predict Chla than nutrient loadings (LIU *et al.*, 2019), which means that part of the explanatory variables could not be able to explain Chla in our model.

8.6.2 Variable Importance

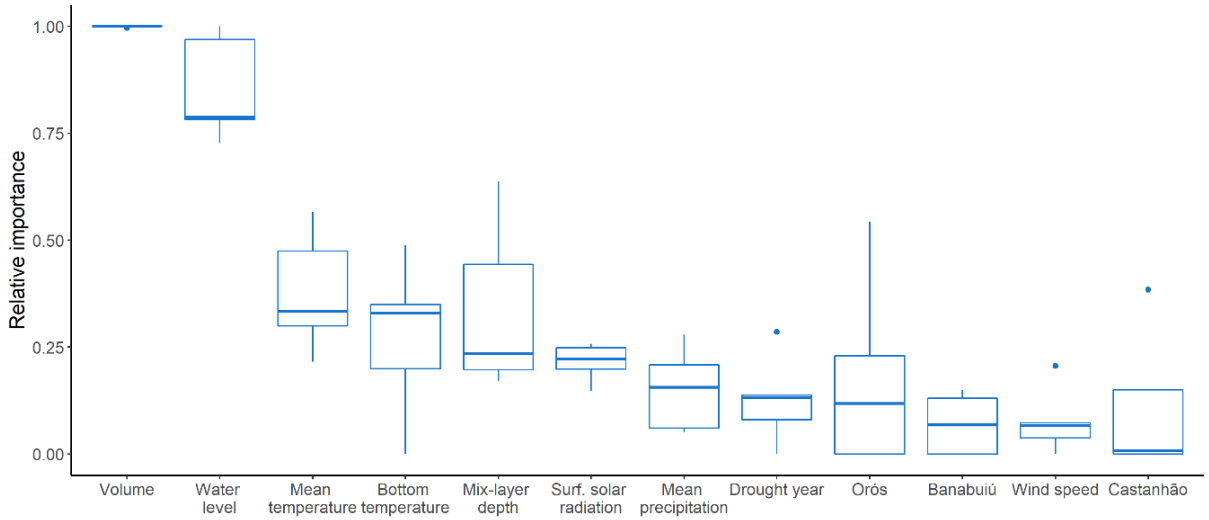
To measure the relative influence of the model's explanatory variables, the importance measure attributed by each predictive model was extracted and scaled using min-max normalization (Figure 45). This approach has been widely used to make machine learning models more interpretable (HASTIE *et al.*, 2009), and can be more accurate than looking only at the correlation between explanatory and dependent variables. Correlation criteria or the goodness of fitness of a linear model are simple and direct strategies to obtain information about a set of variables, but it ignores multicollinearity and interactions between them. Although this study was not intended to perform variable selection, some of the models used here have built-in processes to select the most relevant predictions, such as RF and regularized GLM, the so-called embedded methods (GUYON; ELISSEEFF, 2003).

Radial and kNN models were excluded from this analysis since they do not have a direct importance measure. For RF, GBM, and the regression tree models, the importance corresponds to the reduction in predictive performance obtained by removing the variable from the model. In GLM and MLP, the importance is associated with the weights attributed to each variable.

The boxplots in Figure 45 reveal that water volume was considered the most important predictive variable in all models. The models do not agree regarding the mix-layer depth and bottom temperature importance, as these presented a high variation among them. The dummy variables related to the spatial location of the reservoirs (Castanhão, Orós e Banabuiú) did not seem to significantly influence Chla, indicating that spatial variability could be less important than climate variability, or yet, that the relationships between explanatory variables and Chla are similar for all three reservoirs.

The relative influence of the variables depends on the interactions identified by each model and the procedure used to do it. For example, decision trees choose the optimal variable in

Figure 45 – The relative importance of explanatory variables considering the importance measures of each predictive model, ordered by the median value. Relative importance was scaled between 0 and 1.



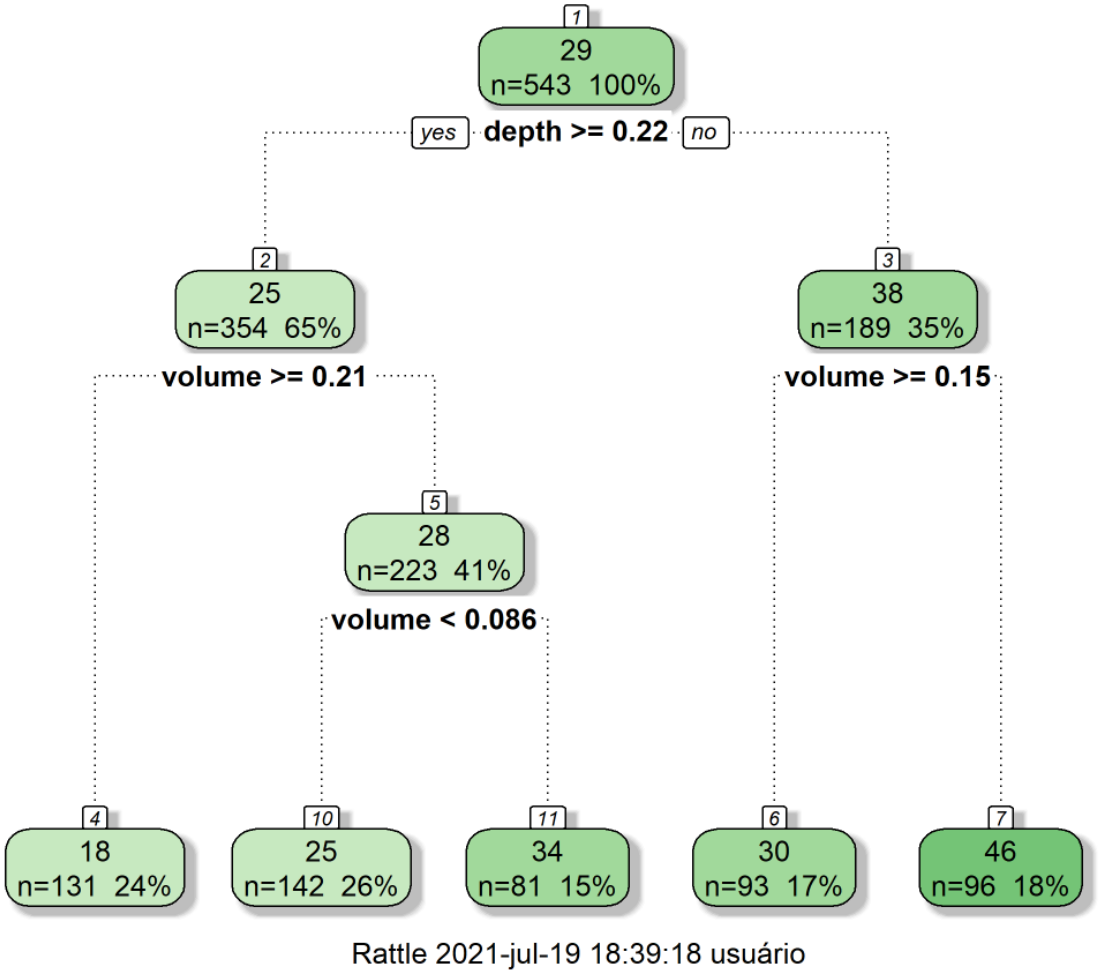
Source: The author.

each split based on the information gained by adding it to the tree. The regression tree constructed to predict Chla had only the mix-layer depth and water volume as predictors (Figure 46). This means that these two variables provide enough information to give us an approximate estimation of Chla. The regression tree alone can be considered a weak predictor, as it is very sensitive to small changes in the dataset and can easily overfit. Since they assume all variables have some interaction between them, it suits well our problem, but it fails to provide accurate estimations of Chla (here, it presented an R^2 of only 0.32). However, it can still give us interesting information on variable importance.

GBM and RF, as explained in the Methods section, combine several regression trees to provide stronger predictive models. RF performs variable selection during its model building process, as the variables used to construct each tree in the ensemble are selected from a random subset of the explanatory variables. The trees are fitted to bootstrap samples of the data, and the importance measure is calculated on the left-out observations (out-of-bag set). The advantage of RF's strategy to calculate variable importance is that it considers both the individual effect and the interactions between the variables (STROBL *et al.*, 2007). GBM, on the other hand, calculates importance on the entire training set instead of using the out-of-bag sets.

To verify the effect of the season on the relationships between the explanatory variables and Chla, all the models were run again for the wet season (observations registered between February and May), and the dry season (observations from the remaining months). Variable importance was extracted for each model and normalized so one could visualize their

Figure 46 – Graphical representation of the regression tree model. The numbers on top of each box (representing a node) are the predicted values of Chla, while n is the number of observations in each node and the number in the bottom right is the percentage of observations in each node. The values of water volume and mix-layer depth are normalized. The variable depth refers to the mix-layer depth and volume is the reservoir water volume.

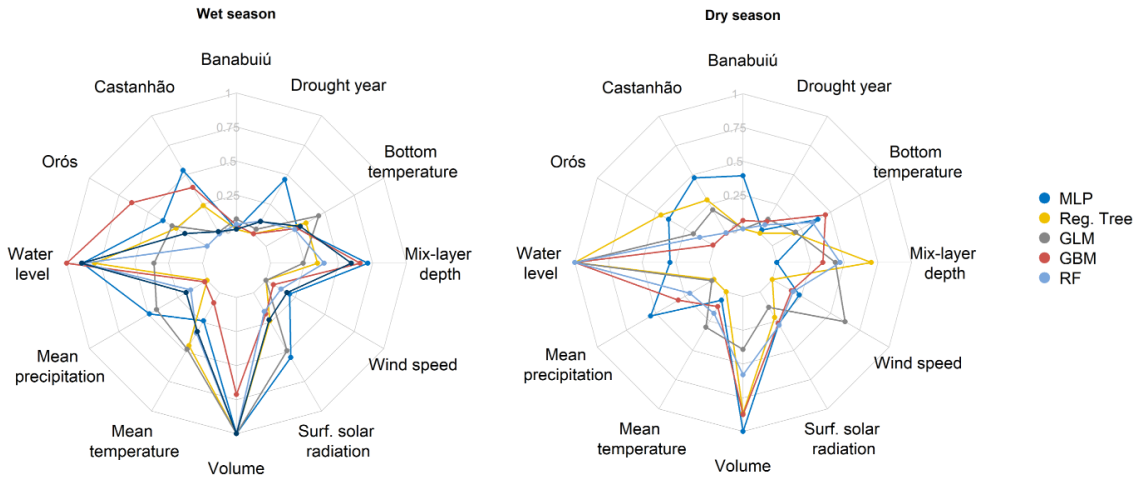


Source: The author.

relative influence on Chla prediction (Figure 47).

Water volume and water level continue to be the most relevant indicators of Chla in both scenarios. However, mix-layer depth and mean temperature seem to be more important in the wet season. It is important to keep in mind that the dry season model has a smaller dataset than the wet season, as it corresponds to the observations of four months only. For this reason, the model can be biased, and more data could be necessary to provide reliable predictions.

Figure 47 – Relative importance of explanatory variables considering separated models for the wet season and dry season.



Source: The author.

8.6.3 Relative influence of hydrological and climate variables on Chla

The PDPs in Figure 48 illustrate the relationships between hydrological and climate variables and Chla. The RF model was selected for this analysis, as it presented the best performance according to all the metrics evaluated. These plots, however, should be interpreted with caution, as they may not display all interactions of the explanatory variables.

Confirming the findings of previous studies, Chla tends to increase as water volume reduces (JUNIOR *et al.*, 2018a; WIEGAND *et al.*, 2021). The decrease in water volume due to evaporation loss, water withdrawals, and extended drought periods are usually associated with higher phosphorus loads in tropical reservoirs (RAULINO *et al.*, 2021; ROCHA; LIMA NETO, 2021b). During the dry period, sediment release and nutrient resuspension are important mechanisms associated with Chla in these reservoirs. Although the effect of internal loading has been pointed as more significant in shallow reservoirs, in the semiarid, precipitation levels come close to zero and inflow decreases drastically during the dry season, so that deep reservoirs reach very low volumes and almost no external loads are carried to them (ROCHA; LIMA NETO, 2022b).

Wind speed did not seem to play an important role in Chla levels, which might be due to reservoirs' morphology and the temporal scale considered here. In deep reservoirs, wind speed is indeed unimportant to Chla, as it is not a relevant driver of water column mixing. Shallow reservoirs, on the other hand, present a significant correlation with nutrient resuspension (ARAúJO *et al.*, 2019; MESQUITA *et al.*, 2020). Past research has indicated that although wind speed affects the dynamics of algal growth and eutrophication, there is a loss of information on

wind dynamics on a monthly scale (STEFANIDIS *et al.*, 2021).

Mix-layer depth has an inverse relationship with Chla, which is consistent with previous findings (STOCKWELL *et al.*, 2020; STEFANIDIS *et al.*, 2021). There are several factors to consider when interpreting this relationship, such as water temperature, reservoir morphology, and the ratio between the mix-layer depth and thermocline depth. In deep reservoirs, stratification is more likely to occur and lake stability tends to increase, with a higher possibility of solute accumulation in the hypolimnion, dissolved oxygen depletion, and phosphorus release from sediments (BUTCHER *et al.*, 2015; KRAEMER *et al.*, 2015; MOURA *et al.*, 2020). But an increase in mix-layer depth also results in a reduction of the light available to phytoplankton (STOCKWELL *et al.*, 2020) and in lower water temperatures, which could inhibit Chla growth (ZHAO *et al.*, 2020).

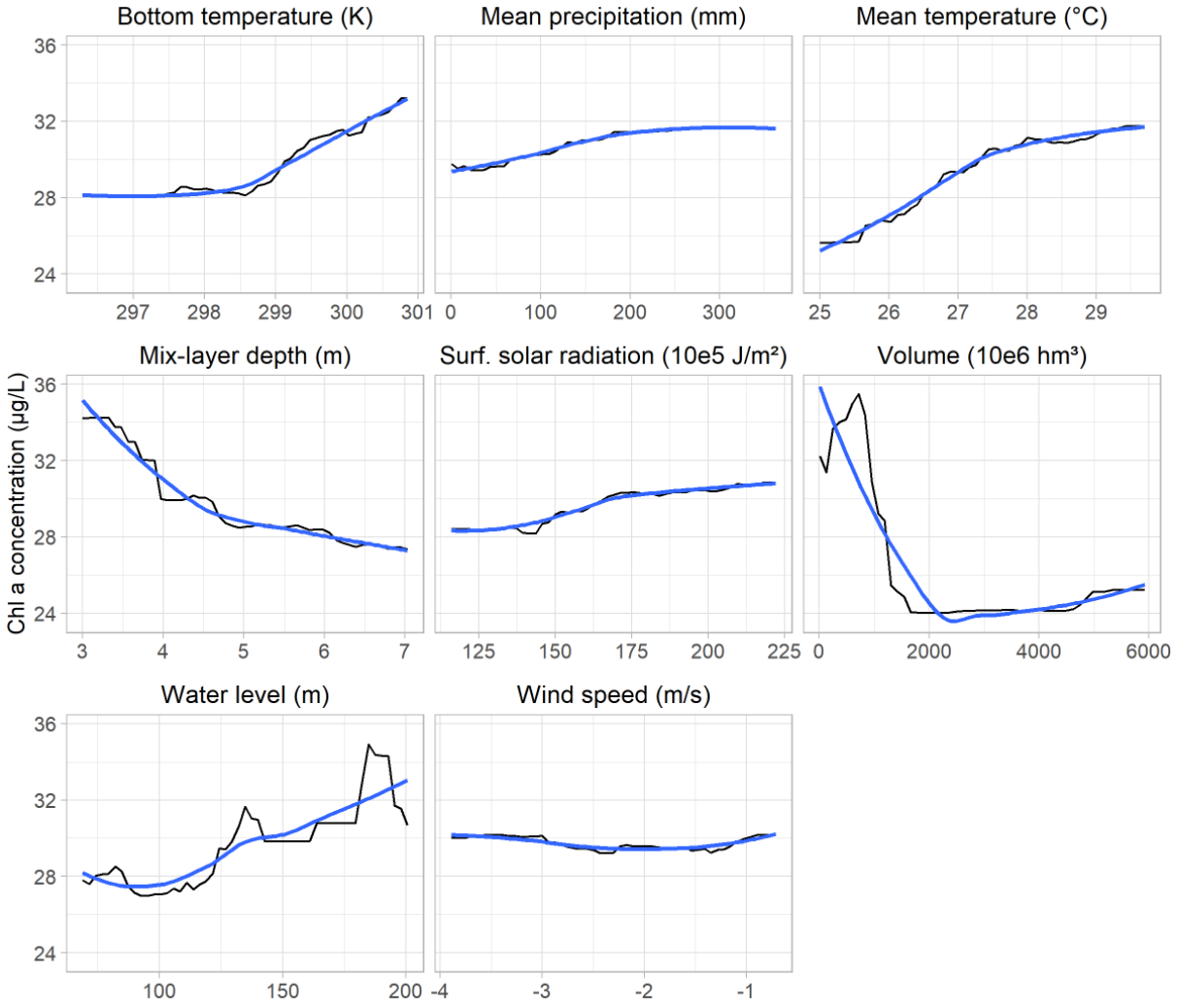
Bottom temperature, mean temperature, solar radiation, and water level have direct relationships with Chla. The first three variables are directly related to each other, and their increase usually enhances phytoplankton productivity (LIU *et al.*, 2019). The direct influence of water level on Chla is surprising, as previous studies have reported the opposite relationship (MEDEIROS *et al.*, 2015; WIEGAND *et al.*, 2020; BRAGA; BECKER, 2020). These studies, however, were performed for small reservoirs, where the relationship between P and Chla is stronger than that for larger reservoirs, i.e., the mechanisms associated with Chla growth are less complex.

The effect of increasing water levels on Chla depends on the quality of the inflow, whether it is related or not to a reduction in the outflow (BAKKER; HILT, 2015), the depth, and the trophic state of the reservoir (COSTA *et al.*, 2015). When precipitation occurs (and water levels start to rise), external loads from rivers and surface runoff add up to internal loads due to thermal stratification and phosphorus release from sediment, which is highly correlated with Chla growth (MOURA *et al.*, 2020). Agriculture and cattle raising are important activities in all reservoirs analyzed here and are the main cause of nonpoint source pollution that increases external total phosphorus loading (ROCHA; LIMA NETO, 2021a).

Although volume and water level are directly related, they have a nonlinear relationship, which can be approximated as a logarithmic curve. Hence, for a certain range, water level fluctuations have little effect on water volume. In this case, Chla growth could be related to some of the factors mentioned above (e.g., the quality of external loads). Reservoir's morphology should also be considered, as the storage depends on the water height-area relationship. Hence,

the effect of water level on Chla might depend on how much water is already stored in the reservoir (i.e., at which position in the water height-area-volume curve the reservoir is), the reservoir's morphology, and the quality of external loads.

Figure 48 – PDPs for predictors of the RF model. The blue smooth line was produced using LOESS to better visualize the relationship between the explanatory and response variables.



Source: The author.

The PDPs for the dry and wet season models were also examined. Except for mean precipitation and wind speed, all variables maintained the patterns observed in the general model. Figure 49 presents the variables with opposing behaviors. While precipitation has a positive effect in the dry season, it presents a negative and almost insignificant effect during the wet season.

One explanation for this behavior is that water volumes tend to be reduced over the dry season. Hence, precipitation can increase nutrient loadings (JEPPESEN *et al.*, 2015; JUNIOR

et al., 2018a) but not have a significant effect on water volume. During the wet season, increased precipitation might induce greater flushing and lower Chla (REICHWALDT; GHADOUANI, 2012). Because the reservoirs have higher water volumes during this season, as the precipitation volume increases, water volume grows exponentially with respect to water level, and Chla might decrease because of mixing and flushing. This effect, however, seems to be not very relevant as produces a little variation on Chla.

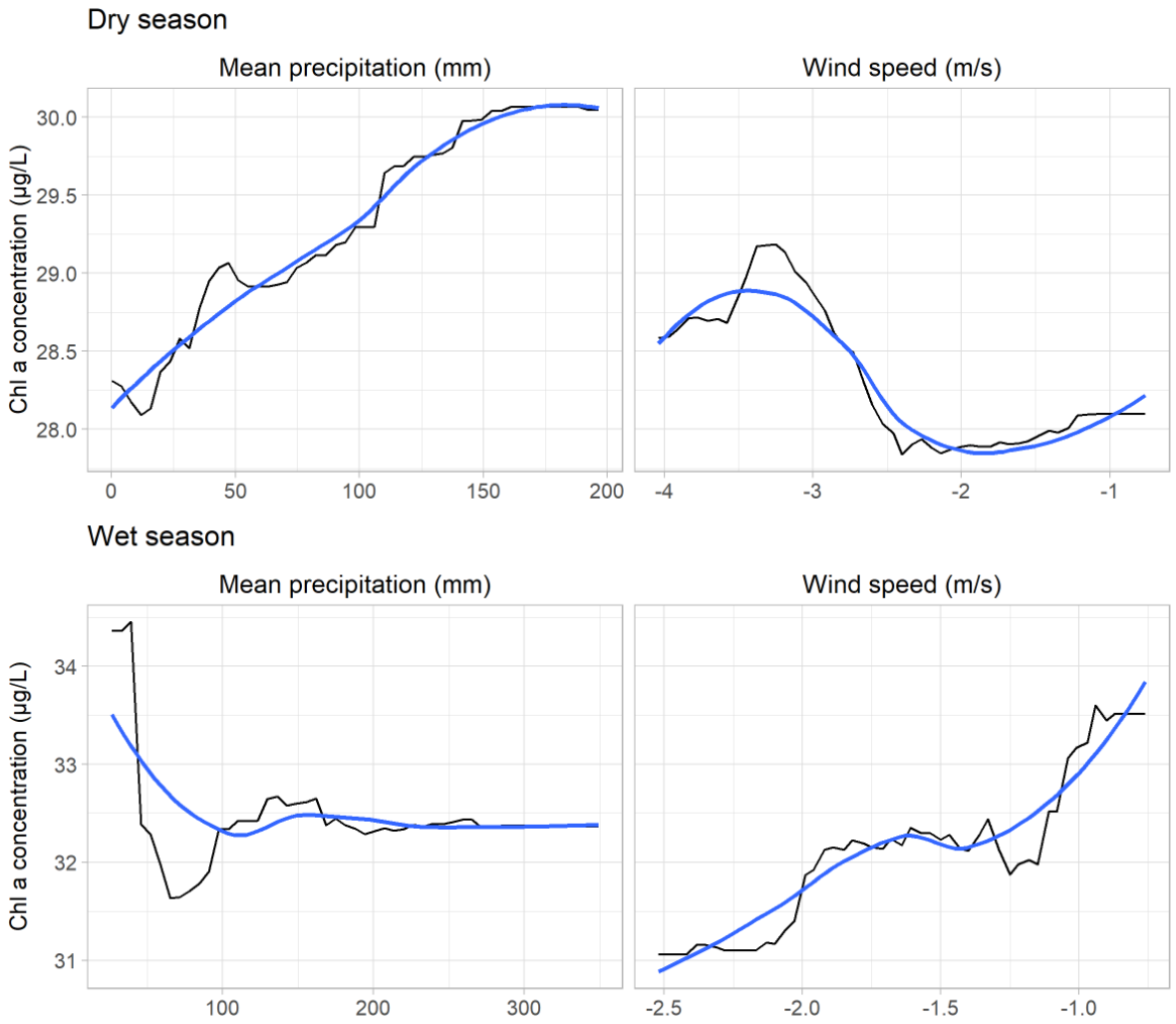
The extent of precipitation influence on Chla is difficult to generalize, as it depends on the intensity and frequency of rainfall events (REICHWALDT; GHADOUANI, 2012; HO; MICHALAK, 2020) and the initial conditions of the reservoir (water volume, trophic state, etc). The reduced stratification during the wet season (LIMA NETO, 2019) can also explain the reduction in Chla during this season, while stronger winds during the dry season can lead to higher Chla concentrations. Hence, precipitation alone is not the only factor to explain Chla fluctuations in both seasons, as its mechanisms are complex.

During the wet season, stronger winds seem to result in a slight decrease of Chla (up to 3 g/L), while in the dry season, it has the opposing effect. The influence of wind speed on Chla can differ according to the water depth, and the sign of this relationship needs further investigation. Previous studies have indicated that increased wind speed can result in greater mixing of the upper layer, thus reducing Chla (STOCKWELL *et al.*, 2020); however, under oligotrophic conditions, stronger winds can carry nutrients to the bottom layer and increase Chla (KAHRU *et al.*, 2010; KIM *et al.*, 2014). This mechanism also depends on the reservoirs' morphology and water level, hence for shallow reservoirs (or for reduced water levels in the dry season), stronger winds can induce resuspension and increase internal nutrient loads (ARAÚJO *et al.*, 2019; ROCHA; LIMA NETO, 2022b). In the wet season, wind-induced resuspension is less significant, as external sources of nutrients play a more important role in Chla fluctuations (ROCHA; LIMA NETO, 2021a).

The relationship between wind speed and internal phosphorus loading has been explored for artificial reservoirs in Ceará, including the ones analyzed here (ROCHA; LIMA NETO, 2022). In this study, the authors found that P release increases with stronger winds (with a threshold value of 3.5 m/s) and the trophic state of the reservoir. As internal loading can increase the risk of eutrophication, wind speed is very likely to be related to Chla in the dry season, when reservoirs become shallower.

PDPs can also be plotted for two variables at the same time (Figure 50). Again, one

Figure 49 – PDPs for precipitation and wind speed for two separate models, one considering the months in the dry season, and the other, the months in the wet season.

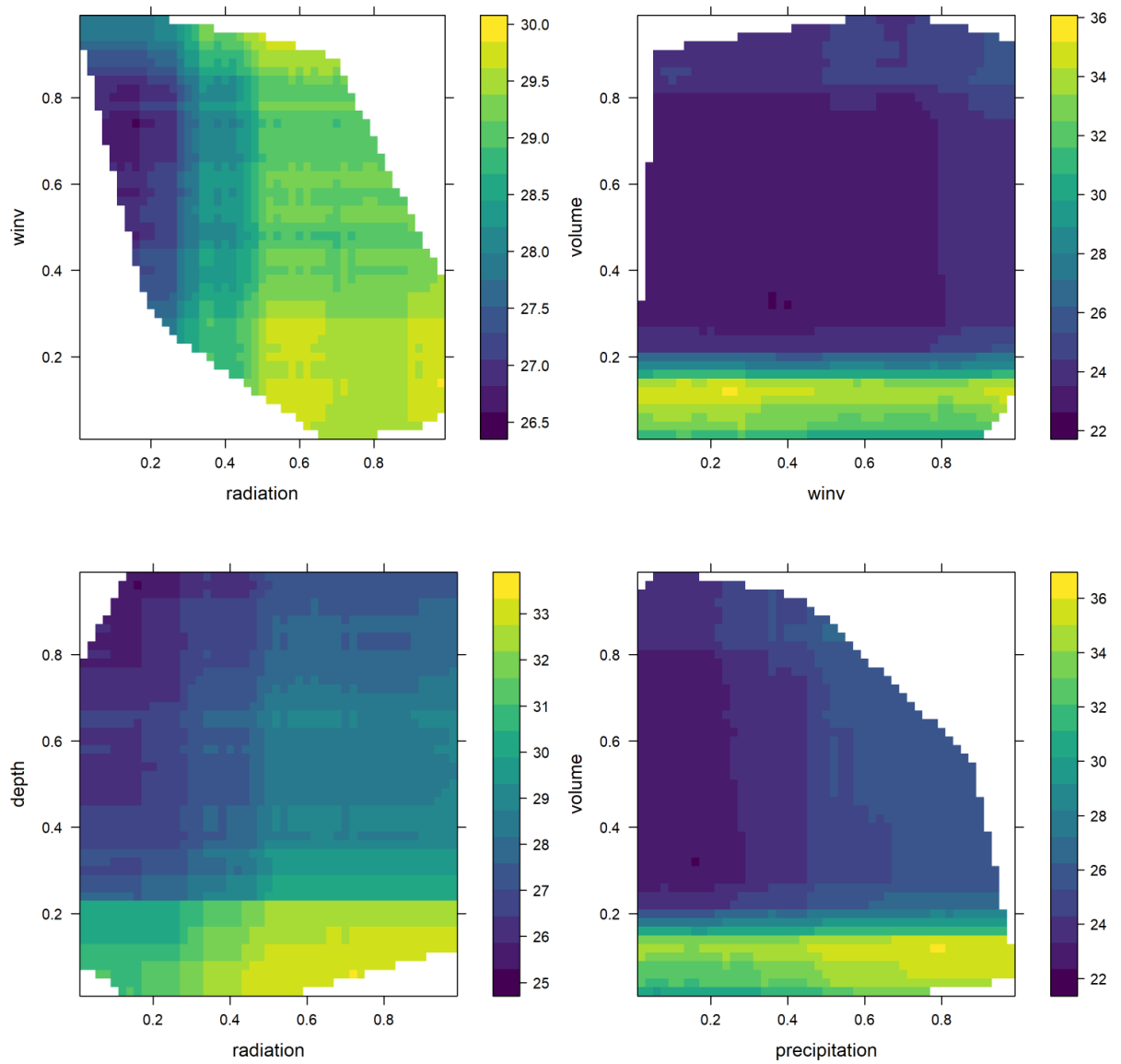


Source: The author.

must be careful when interpreting these plots, as they can show correlations between variables rather than a causal relationship. When considering higher values of solar radiation, wind speed presents an inverse relationship with Chla. Whether the mix-layer is shallow or deep, when solar radiation is higher, Chla tends to increase, a relationship that is confirmed by previous research (BERGER *et al.*, 2006). One can also notice that mix-layer depth seems to have a stronger effect on Chla only up to a certain point.

Wind speed had little effect on Chla when the water volume was constant. Again, this might be related to the size of the reservoirs analyzed here and does not necessarily mean that wind speed does not influence Chla. Previous studies have indicated that wind speed can be an important driver of internal phosphorus loadings in the dry period (ROCHA; LIMA NETO, 2022), thus, this variable should not be neglected.

Figure 50 – PDPs of Chla and the interaction between wind speed (winv), mix-layer depth (depth), solar radiation (radiation), volume, and precipitation. The plots are trimmed to not extrapolate the range of the predictive variables outside the training data. Data were normalized to a range between 0 and 1.



Source: The author.

Precipitation can have distinct effects on nutrient concentrations (HO; MICHALAK, 2020). Our analysis indicates that when the water volume is high, increased precipitation levels mean higher Chla (WIEGAND *et al.*, 2020), while for low water volumes, increased precipitation levels mean lower Chla. This, again, can be related to the climate season, as previously discussed. Although there might have been some information loss due to the temporal resolution of the analysis presented here, the results are consistent with the findings of other studies performed for the semiarid region (MOURA *et al.*, 2020; MESQUITA *et al.*, 2020).

Rather than providing accurate predictions of Chla, the predictive models explored in this study can indicate the magnitude and the overall direction of the relationship between hydro-climatic variables and Chla.

8.7 Conclusion

In the semiarid region, complex mechanisms regulate phytoplankton growth, so that estimates of P may not result in reliable predictions of Chla. This study revealed that a combination of hydrological and climate factors can provide insightful information on Chla fluctuations on a monthly scale. To do that, RF and GBM are the most suitable models, with satisfactory predictive performance.

Looking at the interaction between variables, increasing solar radiation and reducing wind speed result in higher Chla. For a deeper mix-layer, the increase of solar radiation has a positive effect on Chla. Another interesting finding was that precipitation and wind speed present opposing effects on Chla depending on the season. Water level and volume have opposite relationships with Chla: the underlying mechanism associated with Chla is reverted after the dry season (when the internal load is more significant).

These results suggest that climate and hydrological variables have nonlinear relationships with Chla, with an exploratory potential that should not be ignored. Machine learning models can provide important insight on the mechanisms related to Chla increase or decrease in reservoirs, especially when using interpretation methods such as PDPs. By understanding some of the mechanisms associated with hydrological and climatic variability and Chla, policymakers can design more specific strategies to mitigate eutrophication.

There are, however, a few drawbacks of this study, such as the temporal-spatial resolution of the time series, which can hide some of the mechanisms associated with Chla fluctuations. However, extensive field data collection would be needed to overcome this limitation. An interesting approach to be investigated in future studies is the combination of mechanistic water quality modeling and machine learning methods (the so-called scientific machine learning) to assess eutrophication mechanisms. Within this framework, physical, biological and chemical relationships can be incorporated into machine learning modeling, facilitating uncertainty quantification and interpretability.

9 CONCLUSION

The short and long-term planning of a water system involves multiple complexities associated with hydrological, climatic and social processes. The natural sciences are clearly no longer interested in immutable phenomena (PRIGOGINE; STENGERS, 1991), and the announcement of the death of stationarity (MILLY *et al.*, 2008) established the urgent need for water resources management that incorporates the uncertainties of the Earth System.

As the relationships between social and natural environments is not linear and not easily mapped, ML models and statistical learning tools offer a solution to extract knowledge from the available data. Although widely studied in the literature, these tools are unexploited in water resources management. There are also many data sources that have not been sufficiently explored, such as text data (e.g. newspapers, official documents, water resources plans). In this thesis, we develop predictive models and combine existing algorithms with explainable machine learning tools to explore issues associated with water quantity and quality.

We set out investigating the social and economic drivers of residential water demand. We learn that these can explain up to 40% of the low frequency component of water demand, and we might need only a few representative variables to account for this effect - even in coarse spatial scales. The percentage of children in a household, average monthly income and vulnerability to poverty of residents can be used as proxies to estimate water demand patterns in Fortaleza, Ceará. Then, we investigate the seasonal fluctuations of residential water demand and create a predictive model that is able to account for the effect of precipitation and temperature variability on this component. After exploring social, economic and climate effects on water demand, we analyze the implications of price-based demand-side measures. We find that these measures are effective - but it can be unfair to lower income households - while drought awareness is less likely to encourage consumers to save water. During the 2012-2018 drought that severely impacted Ceará, the implementation of a contingency tariff enforced a reduction in water demand of low income households to below the recommended by the World Health Organization (around 100 litres of water per person per day) (KI-MOON; GENERAL, 2010).

We then look at the long-term water supply, and discover that climate variability and the continuous growth in water demand will require alternative water sources to guarantee water supply in the next years. We also find that even by expanding water supply, agricultural water demand might not be fully attended in the next 30 years in Ceará. This means that water resources stakeholders must draw attention to demand-side measures, water losses reduction and

a more refined water accounting system. Finally, we investigate the mechanisms that regulate phytoplankton growth in tropical reservoirs. By using ML, we identify advers relationships between hydrological and climate variables and Chla depending on the season. For example, we conclude that water level and volume have opposite relationships with Chla: the underlying mechanism associated with Chla is reverted after the dry season (when the internal load is more significant).

These studies can serve not only as starting points for discussing new water management policies and adjusting the current ones, but also to provide new models to be deployed and actually incorporated into water supply and management tools (HADJIMICHAEL *et al.*, 2016). However, despite its high predictive power and ability to detect patterns, ML models are heavily conditioned by data availability. Information on income, education, employment and infrastructure at refined spatial scales is not easy to find or is not collected regularly. In addition, diffuse communities, users not served by the water company or with defective meters, or even users of precarious housing, may be absent from the available databases. While the transition to more sophisticated predictive models can mean more detailed information about water users, it can also be a way of reinforcing inequalities and the isolation of some people. Therefore, it is important to take some precautions when applying these models, such as the adoption of data balancing strategies, and to combine its use with social participation.

REFERENCES

ABU-BAKAR, H.; WILLIAMS, L.; HALLETT, S. H. A review of household water demand management and consumption measurement. **Journal of Cleaner Production**, Estados Unidos da América, v. 292, p. 125872, abr. 2021.

ADAMOWSKI, J.; ADAMOWSKI, K.; PROKOPH, A. A Spectral Analysis Based Methodology to Detect Climatological Influences on Daily Urban Water Demand. **Mathematical Geosciences**, Germany, v. 45, n. 1, p. 49–68, jan. 2013.

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**, Estados Unidos da América, v. 15, n. 10, p. 729–743, out. 2010.

ADAMSON, D.; LOCH, A. Possible negative feedbacks from 'gold-plating' irrigation infrastructure. **Agricultural Water Management**, Estados Unidos da América, v. 145, p. 134–144, nov. 2014.

ADEDEJI, K. B.; HAMAM, Y.; ABE, B. T.; ABU-MAHFOUZ, A. M. Leakage Detection and Estimation Algorithm for Loss Reduction in Water Piping Networks. **Water**, Suíça, v. 9, n. 10, p. 773, out. 2017.

AHMED, A. N.; OTHMAN, F. B.; AFAN, H. A.; IBRAHIM, R. K.; FAI, C. M.; HOSSAIN, M. S.; EHTERAM, M.; ELSHAFIE, A. Machine learning methods for better water quality prediction. **Journal of Hydrology**, Estados Unidos da América, v. 578, p. 124084, nov. 2019.

ALI, M.; PRASAD, R.; XIANG, Y.; YASEEN, Z. M. Complete ensemble empirical mode decomposition hybridized with random forest and kernel ridge regression model for monthly rainfall forecasts. **Journal of Hydrology**, Estados Unidos da América, v. 584, p. 124647, maio 2020.

ALTMAN, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. **The American Statistician**, Reino Unido, v. 46, n. 3, p. 175, ago. 1992.

ALTUNKAYNAK, A.; NIGUSSIE, T. A. Monthly Water Consumption Prediction Using Season Algorithm and Wavelet Transform–Based Models. **Journal of Water Resources Planning and Management**, Estados Unidos da América, v. 143, n. 6, p. 04017011, jun. 2017.

ANDRÉ, D. M.; CARVALHO, J. R. Spatial Determinants of Urban Residential Water Demand in Fortaleza, Brazil. **Water Resources Management**, Holanda, v. 28, n. 9, p. 2401–2414, jul. 2014.

APLEY, D. **ALEPlot**: Accumulated Local Effects (ALE) Plots and Partial Dependence (PD) Plots. 2018. Disponível em: https://CRAN.R-project.org/package=ALEPlot. Acesso em: 10 novembro 2020.

ARAÚJO, G. M.; LIMA NETO, I. E.; BECKER, H. Phosphorus dynamics in a highly polluted urban drainage channel shallow reservoir system in the Brazilian semiarid. Anais da **Academia Brasileira de Ciências**, Brasil, v. 91, n. 3, 2019.

ARBUÉS, F.; BARBERÁN, R. Tariffs for Urban Water Services in Spain: Household Size and Equity. **International Journal of Water Resources Development**, Reino Unido, v. 28, n. 1, p. 123–140, mar. 2012.

ARBUÉS, F.; BARBERÁN, R.; VILLANÚA, I. Price impact on urban residential water demand: A dynamic panel data approach. **Water Resources Research**, Estados Unidos da América, v. 40, n. 11, nov. 2004.

AWAD, M.; KHANNA, R. **Efficient Learning Machines:** Theories, Concepts, and Applications for Engineers and System Designers. Berkeley, CA: Apress, 2015.

BACHE, S. M.; WICKHAM, H. magrittr: A Forward-Pipe Operator for R. 2020. Disponível em: https://CRAN.R-project.org/package=magrittr.

BAKKER, E. S.; HILT, S. Impact of water-level fluctuations on cyanobacterial blooms: options for management. **Aquatic Ecology**, Holanda, v. 50, n. 3, p. 485–498, dez. 2015.

BARRETO, F. A. F. D.; MENEZES, A. S. B. **Desenvolvimento econômico do Ceará**: Evidências recentes e reflexões. Fortaleza: IPECE, 2014. Disponível em: https://www.ipece.ce.gov.br/wp-content/uploads/sites/45/2015/02/Desenvolvimento_Economico_do_Ceara_Evidencias_Recentes_e_Reflexoes.pdf. Acesso em: 21 jan. 2021.

BATA, M. H.; CARRIVEAU, R.; TING, D. S.-K. Short-Term Water Demand Forecasting Using Nonlinear Autoregressive Artificial Neural Networks. **Journal of Water Resources Planning and Management**, Estados Unidos da América, v. 146, n. 3, p. 04020008, mar. 2020. ISSN 0733-9496, 1943-5452.

BDOUR, A. N.; HAMDI, M. R.; TARAWNEH, Z. Perspectives on sustainable wastewater treatment technologies and reuse options in the urban areas of the Mediterranean region. **Desalination**, Holanda, v. 237, n. 1-3, p. 162–174, fev. 2009.

BENNETT, C.; STEWART, R. A.; BEAL, C. D. ANN-based residential water end-use demand forecasting model. **Expert Systems with Applications**, Reino Unido, v. 40, n. 4, p. 1014–1023, mar. 2013.

BERBEL, J.; GUTIÉRREZ-MARTíN, C.; EXPÓSITO, A. Impacts of irrigation efficiency improvement on water use, water consumption and response to water price at field level. **Agricultural Water Management**, Holanda, v. 203, p. 423–429, abr. 2018.

BERGER, S. A.; DIEHL, S.; STIBOR, H.; TROMMER, G.; RUHENSTROTH, M.; WILD, A.; WEIGERT, A.; JäGER, C. G.; STRIEBEL, M. Water temperature and mixing depth affect timing and magnitude of events during spring succession of the plankton. **Oecologia**, Alemanha, v. 150, n. 4, p. 643–654, dez. 2006.

BERGMEIR, C.; BENÍTEZ, J. M. Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. **Journal of Statistical Software**, Estados Unidos, v. 46, n. 7, p. 1–26, 2012.

BERNOULLI, D. Exposition of a New Theory on the Measurement of Risk. **Econometrica**, Reino Unido, v. 22, n. 1, p. 23, jan. 1954.

BEZANSON, J.; EDELMAN, A.; KARPINSKI, S.; SHAH, V. B. Julia: A Fresh Approach to Numerical Computing. **SIAM Review**, Estados Unidos da América, v. 59, n. 1, p. 65–98, jan. 2017.

BISHOP, C. M. **Neural networks for pattern recognition**. [*S. l.*]: Oxford University Press, 1995.

BLAUW, A. N.; BENINCà, E.; LAANE, R. W.; GREENWOOD, N.; HUISMAN, J. Predictability and environmental drivers of chlorophyll fluctuations vary across different time scales and regions of the North Sea. **Progress in Oceanography**, Reino Unido, v. 161, p. 1–18, fev. 2018.

BOLORINOS, J.; AJAMI, N. K.; RAJAGOPAL, R. Consumption Change Detection for Urban Planning: Monitoring and Segmenting Water Customers During Drought. **Water Resources Research**, Estados Unidos da América, v. 56, n. 3, mar. 2020.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. *In:* WORKSHOP ON COMPUTATIONAL LEARNING THEORY, 5., 1992, Pennsylvania. **Proceedings** [...]. Pennsylvania, USA: ACM, 1992. p. 144–152. Acesso em: 17 fev. 2021.

BRAGA, B.; KELMAN, J. Facing the challenge of extreme climate: the case of Metropolitan Sao Paulo. **International Journal of Water Resources Development**, Reino Unido, v. 36, n. 2-3, p. 278–291, mar. 2020.

BRAGA, G. G.; BECKER, V. Influence of water volume reduction on the phytoplankton dynamics in a semi-arid man-made lake: A comparison of two morphofunctional approaches. **Anais da Academia Brasileira de Ciências**, Brasil, v. 92, n. 1, p. e20181102, 2020.

BREIMAN, L. Random Forests. Machine Learning, v. 45, n. 1, p. 5–32, 2001.

BRENTAN, B. M.; JR., E. L.; HERRERA, M.; IZQUIERDO, J.; PÉREZ-GARCíA, R. Hybrid regression model for near real-time urban water demand forecasting. **Journal of Computational and Applied Mathematics**, Holanda, v. 309, p. 532–541, jan. 2017.

BRITO, M. M.; KUHLICKE, C.; MARX, A. Near-real-time drought impact assessment: a text mining approach on the 2018/19 drought in Germany. **Environmental Research Letters**, Reino Unido, v. 15, n. 10, p. 1040a9, out. 2020.

BUTCHER, J. B.; NOVER, D.; JOHNSON, T. E.; CLARK, C. M. Sensitivity of lake thermal and mixing dynamics to climate change. **Climatic Change**, Holanda, v. 129, n. 1, p. 295–305, jan. 2015.

CAMPOS, J.; SOUZA FILHO, F. A.; LIMA, H. V. C. Risks and uncertainties in reservoir yield in highly variable intermittent rivers: case of the Castanhão Reservoir in semi-arid Brazil. **Hydrological Sciences Journal**, Reino Unido, v. 59, n. 6, p. 1184–1195, jun. 2014.

CAMPOS, J. N. B. Paradigms and Public Policies on Drought in Northeast Brazil: A Historical Perspective. **Environmental Management**, Estados Unidos da América, v. 55, n. 5, p. 1052–1063, maio 2015.

CARDELL-OLIVER, R.; WANG, J.; GIGNEY, H. Smart Meter Analytics to Pinpoint Opportunities for Reducing Household Water Use. **Journal of Water Resources Planning and Management**, Estados Unidos da América, v. 142, n. 6, p. 04016007, jun. 2016.

CEARÁ. Lei n° 11.996, de 24 de julho de 1992. Dispõe sobre a Política Estadual de Recursos Hídricos, institui o Sistema Integrado de gestão de Recursos Hídricos e dá outras providências. **Diário oficial do Estado**, Ceará, p. 21, 29 jul. 1992. Disponível em: http://www.srh.ce.gov.br/leis/images/lei11996.pdf, Acesso em: 20 mar. 2021.

CHANG, H.; BONNETTE, M. R.; STOKER, P.; CROW-MILLER, B.; WENTZ, E. Determinants of single family residential water use across scales in four western US cities. **Science of The Total Environment**, Holanda, v. 596-597, p. 451–464, out. 2017.

CHANG, H.; PARANDVASH, G. H.; SHANDAS, V. Spatial Variations of Single-Family Residential Water Consumption in Portland, Oregon. **Urban Geography**, Reino Unido, v. 31, n. 7, p. 953–972, out. 2010. ISSN 0272-3638, 1938-2847.

CHANG, H.; PRASKIEVICZ, S.; PARANDVASH, H. Sensitivity of Urban Water Consumption to Weather and Climate Variability at Multiple Temporal Scales: The Case of Portland, Oregon. **International Journal of Geospatial and Environmental Research**, Estados Unidos da América, v.1, n. 1, 2014.

CHAUDHARY, V.; BHATIA, R.; AHLAWAT, A. K. A novel Self-Organizing Map (SOM) learning algorithm with nearest and farthest neurons. **Alexandria Engineering Journal**, Egito, v. 53, n. 4, p. 827–831, dez. 2014.

CHEN, G.; LONG, T.; XIONG, J.; BAI, Y. Multiple Random Forests Modelling for Urban Water Consumption Forecasting. **Water Resources Management**, Holanda, v. 31, n. 15, p. 4715–4729, dez. 2017.

CHU, L.; GRAFTON, R. Q. Short-term Pain for Long-term Gain: Urban Water Pricing and the Risk-adjusted User Cost. **Water Economics and Policy**, Singapura, v. 05, n. 02, p. 1871005, abr. 2019.

CISNEROS, B. E. J.; OKI, T.; ARNELL, N. W.; BENITO, G.; COGLEY, J. G.; DöLL, P.; JIANG, T.; MWAKALILA, S. S. **Freshwater resources**. *In:* Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. [S. l.]: Cambridge University Press, 2014. p. 229–269.

CLEVELAND, R. B.; CLEVELAND, W. S.; TERPENNING, I. STL: A Seasonal-Trend Decomposition Procedure Based on Loess. **Journal of Official Statistics**, Suíça, v. 6, n. 1, p. 3–73, 1990.

COMINOLA, A.; GIULIANI, M.; PIGA, D.; CASTELLETTI, A.; RIZZOLI, A. Benefits and challenges of using smart meters for advancing residential water demand modeling and management: A review. **Environmental Modelling & Software**, Holanda, v. 72, p. 198–214, out. 2015.

COMINOLA, A.; NGUYEN, K.; GIULIANI, M.; STEWART, R. A.; MAIER, H. R.; CASTELLETTI, A. Data Mining to Uncover Heterogeneous Water Use Behaviors From Smart Meter Data. **Water Resources Research**, Estados Unidos da América, v. 55, n. 11, p. 9315–9333, nov. 2019.

COMINOLA, A.; SPANG, E.; GIULIANI, M.; CASTELLETTI, A.; LUND, J.; LOGE, F. Segmentation analysis of residential water-electricity demand for customized demand-side management programs. **Journal of Cleaner Production**, Reino Unido, v. 172, p. 1607–1619, jan. 2018.

COSGROVE, W. J.; LOUCKS, D. P. Water management: Current and future challenges and research directions. **Water Resources Research**, Estados Unidos da América, v. 51, n. 6, p. 4823–4839, 2015.

COSTA, M. R. A. d.; ATTAYDE, J. L.; BECKER, V. Effects of water level reduction on the dynamics of phytoplankton functional groups in tropical semi-arid shallow lakes. **Hydrobiologia**, Holanda, v. 778, n. 1, p. 75–89, dez. 2015.

DALHUISEN, J. M.; FLORAX, R. J. G. M.; GROOT, H. L. F. de; NIJKAMP, P. Price and Income Elasticities of Residential Water Demand: A Meta-Analysis. **Land Economics**, Estados Unidos da América, v. 79, n. 2, p. 292–308, maio 2003.

DEYÀ-TORTELLA, B.; GARCIA, C.; NILSSON, W.; TIRADO, D. The effect of the water tariff structures on the water consumption in Mallorcan hotels: Water tariff structures and hotel water consumption. **Water Resources Research**, Estados Unidos da América, v. 52, n. 8, p. 6386–6403, ago. 2016.

DIAS, T. F.; KALBUSCH, A.; HENNING, E. Factors influencing water consumption in buildings in southern Brazil. **Journal of Cleaner Production**, Reino Unido, v. 184, p. 160–167, maio 2018.

DILLON, P. J.; RIGLER, F. H. The phosphorus-chlorophyll relationship in lakes1,2: Phosphorus-chlorophyll relationship. **Limnology and Oceanography**, Estados Unidos da América, v. 19, n. 5, p. 767–773, set. 1974.

DOWSON, O.; KAPELEVICH, L. SDDP.jl: A Julia Package for Stochastic Dual Dynamic Programming. **INFORMS Journal on Computing**, v. 33, n. 1, p. 27–33, 2021.

DRAGOMIRETSKIY, K.; ZOSSO, D. Variational Mode Decomposition. **IEEE Transactions on Signal Processing**, Estados Unidos da América, v. 62, n. 3, p. 531–544, fev. 2014.

DUERR, I.; MERRILL, H. R.; WANG, C.; BAI, R.; BOYER, M.; DUKES, M. D.; BLIZNYUK, N. Forecasting urban household water demand with statistical and machine learning methods using large space-time data: A Comparative study. **Environmental Modelling & Software**, Holanda, v. 102, p. 29–38, abr. 2018.

DUNN, J. C. Well-Separated Clusters and Optimal Fuzzy Partitions. **Journal of Cybernetics**, Reino Unido, v. 4, n. 1, p. 95–104, jan. 1974.

DUNSTAN, P. K.; FOSTER, S. D.; KING, E.; RISBEY, J.; O'KANE, T. J.; MONSELESAN, D.; HOBDAY, A. J.; HARTOG, J. R.; THOMPSON, P. A. Global patterns of change and variation in sea surface temperature and chlorophyll a. **Scientific Reports**, Reino Unido, v. 8, n. 1, p. 14624, out. 2018.

EASTERLING, D. R.; MEEHL, G. A.; PARMESAN, C.; CHANGNON, S. A.; KARL, T. R.; MEARNS, L. O. Climate Extremes: Observations, Modeling, and Impacts. **Science**, Estados Unidos da América, v. 289, n. 5487, p. 2068–2074, set. 2000.

ESPEY, M.; ESPEY, J.; SHAW, W. D. Price elasticity of residential demand for water: A meta-analysis. **Water Resources Research**, Estados Unidos da América, v. 33, n. 6, p. 1369–1374, jun. 1997.

FENG, Z.-k.; NIU, W.-j.; TANG, Z.-y.; JIANG, Z.-q.; XU, Y.; LIU, Y.; ZHANG, H.-r. Monthly runoff time series prediction by variational mode decomposition and support vector machine based on quantum-behaved particle swarm optimization. **Journal of Hydrology**, Estados Unidos da América, v. 583, p. 124627, abr. 2020.

FICKLIN, D. L.; NULL, S. E.; ABATZOGLOU, J. T.; NOVICK, K. A.; MYERS, D. T. Hydrological Intensification Will Increase the Complexity of Water Resource Management. **Earth's Future**, Estados Unidos da América, v. 10, n. 3, p. e2021EF002487, 2022.

FIORILLO, D.; KAPELAN, Z.; XENOCHRISTOU, M.; PAOLA, F. D.; GIUGNI, M. Assessing the Impact of Climate Change on Future Water Demand using Weather Data. **Water Resources Management**, Holanda, v. 35, n. 5, p. 1449–1462, mar. 2021.

FIRAT, M.; YURDUSEV, M. A.; TURAN, M. E. Evaluation of Artificial Neural Network Techniques for Municipal Water Consumption Modeling. **Water Resources Management**, Holanda, v. 23, n. 4, p. 617–632, mar. 2009.

FRAGA, C. C.; MEDELLÍN-AZUARA, J.; MARQUES, G. F. Planning for infrastructure capacity expansion of urban water supply portfolios with an integrated simulation-optimization approach. **Sustainable Cities and Society**, Holanda, v. 29, p. 247–256, fev. 2017.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, Estados Unidos da América, v. 29, n. 5, out. 2001.

FRUTAL. **Plano Diretor de Agricultura Irrigada do Ceará**. Ceará: ADECE, 2013. Disponível em: http://www.adece.ce.gov.br/downloads/plano-diretor-de-agricultura-irrigada-do-ceara, Acesso em: 17 nov. 2022.

GALELLI, S.; CASTELLETTI, A. Tree-based iterative input variable selection for hydrological modeling: Tree-Based Input Selection. **Water Resources Research**, Estados Unidos da América, v. 49, n. 7, p. 4295–4310, jul. 2013.

GAO, Y.; GE, G.; SHENG, Z.; SANG, E. Analysis and Solution to the Mode Mixing Phenomenon in EMD. *In:* CONGRESS ON IMAGE AND SIGNAL PROCESSING. 5., 2008, Sanya. **Proceedings [...]**. Sanya, China: IEEE, 2008. p. 223–227.

GARCIA, J.; SALFER, L. R.; KALBUSCH, A.; HENNING, E. Identifying the Drivers of Water Consumption in Single-Family Households in Joinville, Southern Brazil. **Water**, Suíça, v. 11, n. 10, p. 1990, set. 2019.

GARCIA, X.; PARGAMENT, D. Reusing wastewater to cope with water scarcity: Economic, social and environmental considerations for decision-making. **Resources, Conservation and Recycling**, Holanda, v. 101, p. 154–166, ago. 2015.

GARCÍA-RUBIO, M.; RUIZ-VILLAVERDE, A.; GONZÁLEZ-GÓMEZ, F. Urban Water Tariffs in Spain: What Needs to Be Done? **Water**, Suíça, v. 7, n. 12, p. 1456–1479, mar. 2015.

GARRONE, P.; GRILLI, L.; MARZANO, R. Price elasticity of water demand considering scarcity and attitudes. **Utilities Policy**, Reino Unido, v. 59, p. 100927, ago. 2019.

GENUER, R.; POGGI, J.-M.; TULEAU-MALOT, C. Variable selection using random forests. **Pattern Recognition Letters**, Holanda, v. 31, n. 14, p. 2225–2236, out. 2010.

GHARABAGHI, S.; STAHL, E.; BONAKDARI, H. Integrated nonlinear daily water demand forecast model. **Journal of Hydrology**, Estados Unidos da América, v. 579, p. 124182, dez. 2019.

GHOLIZADEH, M.; MELESSE, A.; REDDI, L. A Comprehensive Review on Water Quality Parameters Estimation Using Remote Sensing Techniques. **Sensors**, Suíça, v. 16, n. 8, p. 1298, ago. 2016.

GONZALES, P.; AJAMI, N. Social and Structural Patterns of Drought-Related Water Conservation and Rebound. **Water Resources Research,** Estados Unidos da América, v. 53, n. 12, p. 10619–10634, dez. 2017.

GRAFTON, R. Q.; CHU, L.; KOMPAS, T.; WARD, M. Volumetric water pricing, social surplus and supply augmentation. **Water Resources and Economics**, Holanda, v. 6, p. 74–87, jul. 2014.

GRANDE, M. H. D.; GALVÃO, C. D. O.; MIRANDA, L. I. B. D.; SOBRINHO, L. D. G. The perception of users about the impacts of water rationing on their household routines. **Ambiente & Sociedade**, Brasil, v. 19, n. 1, p. 163–182, mar. 2016.

GREENWELL, B.; BOEHMKE, B.; CUNNINGHAM, J.; DEVELOPERS, G. B. M. **gbm:** Generalized Boosted Regression Models. Estados Unidos da América, 2020. Disponível em: https://CRAN.R-project.org/package=gbm. Acesso em: 21 fev. 2022.

GREENWELL, B. M. pdp: An R Package for Constructing Partial Dependence Plots. **The R Journal,** v. 9, n. 1, p. 421–436, 2017. Disponível em: https://journal.r-project.org/archive/2017/RJ-2017-016/index.html. Acesso em: 21 fev. 2022.

GREVE, P.; KAHIL, T.; MOCHIZUKI, J.; SCHINKO, T.; SATOH, Y.; BUREK, P.; FISCHER, G.; TRAMBEREND, S.; BURTSCHER, R.; LANGAN, S.; WADA, Y. Global assessment of water challenges under uncertainty in water scarcity projections. **Nature Sustainability**, Reino Unido, v. 1, n. 9, p. 486–494, set. 2018.

GULIS, G. Life Expectancy as an Indicator of Environmental Health. **European Journal of Epidemiology**, Holanda, v. 16, n. 2, p. 161–165, 2000.

GUO, G.; LIU, S.; WU, Y.; LI, J.; ZHOU, R.; ZHU, X. Short-Term Water Demand Forecast Based on Deep Learning Method. **Journal of Water Resources Planning and Management**, Estados Unidos da América, v. 144, n. 12, p. 04018076, dez. 2018.

GUTIÉRREZ, A. P. A.; ENGLE, N. L.; NYS, E. D.; MOLEJÓN, C.; MARTINS, E. S. Drought preparedness in Brazil. **Weather and Climate Extremes**, Holanda, v. 3, p. 95–106, jun. 2014.

GUYON, I.; ELISSEEFF, A. An Introduction to Variable and Feature Selection. **Journal of Machine Learning Research**, Estados Unidos da América, v. 3, p. 1157–1182, 2003.

HADJIMICHAEL, A.; COMAS, J.; COROMINAS, L. Do machine learning methods used in data mining enhance the potential of decision support systems? A review for the urban water sector. **AI Communications**, Holanda, v. 29, n. 6, p. 747–756, dez. 2016.

HAMILTON, N.; FERRY, M. **vmd: Variational Mode Decomposition**. [*S.* l.], 2017. Disponível em: https://CRAN.R-project.org/package=vmd. Acesso em: 27 set. 2021.

HAQUE, M. M.; RAHMAN, A.; HAGARE, D.; KIBRIA, G. Probabilistic Water Demand Forecasting Using Projected Climatic Data for Blue Mountains Water Supply System in Australia. **Water Resources Management**, Holanda, v. 28, n. 7, p. 1959–1971, maio 2014.

HAQUE, M. M.; SOUZA, A. de; RAHMAN, A. Water Demand Modelling Using Independent Component Regression Technique. **Water Resources Management**, Holanda, v. 31, n. 1, p. 299–312, jan. 2017.

HARRIS, I.; OSBORN, T. J.; JONES, P.; LISTER, D. Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. **Scientific Data**, Reino Unido, v. 7, n. 1, p. 109, abr. 2020.

HASELBECK, V.; KORDILLA, J.; KRAUSE, F.; SAUTER, M. Self-organizing maps for the identification of groundwater salinity sources based on hydrochemical data. **Journal of Hydrology**, Holanda, v. 576, p. 610–619, set. 2019.

HASHIMOTO, T.; STEDINGER, J. R.; LOUCKS, D. P. Reliability, resiliency, and vulnerability criteria for water resource system performance evaluation. **Water Resources Research**, Estados Unidos da América, v. 18, n. 1, p. 14–20, fev. 1982.

HASTENRATH, S.; HELLER, L. Dynamics of climatic hazards in northeast Brazil. **Quarterly Journal of the Royal Meteorological Society**, Estados Unidos da América, v. 103, n. 435, p. 77–92, jan. 1977.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. **The elements of statistical learning**: data mining, inference, and prediction. 2. ed. New York: Springer, 2009.

HAUSWIRTH, S. M.; BIERKENS, M. F.; BEIJK, V.; WANDERS, N. The potential of data driven approaches for quantifying hydrological extremes. **Advances in Water Resources**, Reino Unido, v. 155, p. 104017, set. 2021.

HEMATI, A.; RIPPY, M. A.; GRANT, S. B.; DAVIS, K.; FELDMAN, D. Deconstructing Demand: The Anthropogenic and Climatic Drivers of Urban Water Consumption. **Environmental Science & Technology**, Estados Unidos da América, v. 50, n. 23, p. 12557–12566, dez. 2016.

HENRY, L.; WICKHAM, H. purrr: Functional Programming Tools. 2020. Disponível em: https://CRAN.R-project.org/package=purrr.

HERNÁNDEZ-SANCHO, F.; MOLINOS-SENANTE, M.; SALA-GARRIDO, R. Economic valuation of environmental benefits from wastewater treatment processes: An empirical approach for Spain. **Science of The Total Environment**, Holanda, v. 408, n. 4, p. 953–957, jan. 2010.

HERRERA, M.; TORGO, L.; IZQUIERDO, J.; PéREZ-GARCíA, R. Predictive models for forecasting hourly urban water demand. **Journal of Hydrology**, Estados Unidos da América, v. 387, n. 1-2, p. 141–150, jun. 2010.

HERSBACH, H.; BELL, B.; BERRISFORD, P.; HIRAHARA, S.; HORáNYI, A.; MUñOZ-SABATER, J.; NICOLAS, J.; PEUBEY, C.; RADU, R.; SCHEPERS, D.; SIMMONS, A.; SOCI, C.; ABDALLA, S.; ABELLAN, X.; BALSAMO, G.; BECHTOLD, P.; BIAVATI, G.; BIDLOT, J.; BONAVITA, M.; CHIARA, G.; DAHLGREN, P.; DEE, D.; DIAMANTAKIS, M.; DRAGANI, R.; FLEMMING, J.; FORBES, R.; FUENTES, M.; GEER, A.; HAIMBERGER, L.; HEALY, S.; HOGAN, R. J.; HÓLM, E.; JANISKOVÁ, M.; KEELEY, S.; LALOYAUX, P.; LOPEZ, P.; LUPU, C.; RADNOTI, G.; ROSNAY, P.; ROZUM, I.; VAMBORG, F.; VILLAUME, S.; THÉPAUT, J.-N. The ERA5 global reanalysis. **Quarterly Journal of the Royal Meteorological Society**, Estados Unidos da América, v. 146, n. 730, p. 1999–2049, jul. 2020.

HIRSCH, R. M.; ARCHFIELD, S. A. Not higher but more often. **Nature Climate Change**, Reino Unido, v. 5, n. 3, p. 198–199, mar. 2015.

HO, J. C.; MICHALAK, A. M. Exploring temperature and precipitation impacts on harmful algal blooms across continental U.S. lakes. **Limnology and Oceanography**, Estados Unidos da América, v. 65, n. 5, p. 992–1009, maio 2020.

HOUSE-PETERS, L.; PRATT, B.; CHANG, H. Effects of Urban Spatial Structure, Sociodemographics, and Climate on Residential Water Consumption in Hillsboro, Oregon. **Journal of the American Water Resources Association**, Estados Unidos da América, v. 46, n. 3, p. 461–472, jun. 2010.

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

HUANG, N. E.; SHEN, Z.; LONG, S. R.; WU, M. C.; SHIH, H. H.; ZHENG, Q.; YEN, N.-C.; TUNG, C. C.; LIU, H. H. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. **Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences**, v. 454, n. 1971, p. 903–995, mar. 1998.

HUSSIEN, W. A.; MEMON, F. A.; SAVIC, D. A. Assessing and Modelling the Influence of Household Characteristics on Per Capita Water Consumption. **Water Resources Management**, v. 30, n. 9, p. 2931–2955, jul. 2016.

IIAMES, J. S.; SALLS, W. B.; MEHAFFEY, M. H.; NASH, M. S.; CHRISTENSEN, J. R.; SCHAEFFER, B. A. Modeling Anthropogenic and Environmental Influences on Freshwater Harmful Algal Bloom Development Detected by MERIS Over the Central United States. **Water Resources Research**, Estados Unidos da América, v. 57, n. 10, out. 2021.

IPLANFOR. **Fortaleza 2040**. [*S. l.*], 2015.

JEPPESEN, E.; BRUCET, S.; NASELLI-FLORES, L.; PAPASTERGIADOU, E.; STEFANIDIS, K.; NõGES, T.; NõGES, P.; ATTAYDE, J. L.; ZOHARY, T.; COPPENS, J.; BUCAK, T.; MENEZES, R. F.; FREITAS, F. R. S.; KERNAN, M.; SøNDERGAARD, M.; BEKLIOğLU, M. Ecological impacts of global warming and water abstraction on lakes and reservoirs due to changes in water level and related changes in salinity. **Hydrobiologia**, Holanda, v. 750, n. 1, p. 201–227, maio 2015.

JONES, J.; BACHMANN, B. W. Prediction of phosphorus and chlorophyll levels in lakes. **Water Pollution Control Federation**, Estados Unidos da América, p. 2176–2182, 1976.

JUNIOR, C. A. N. d. R.; COSTA, M. R. A. d.; MENEZES, R. F.; ATTAYDE, J. L.; BECKER, V. Water volume reduction increases eutrophication risk in tropical semi-arid reservoirs. **Acta Limnologica Brasiliensia**, Brasil, v. 30, n. 0, abr. 2018.

JUNIOR, F. d. C. V.; JONES, C.; GANDU, A. W. Interannual and Intraseasonal Variations of the Onset and Demise of the Pre-Wet Season and the Wet Season in the Northern Northeast Brazil. **Revista Brasileira de Meteorologia**, Brasil, v. 33, n. 3, p. 472–484, set. 2018.

KAHRU, M.; GILLE, S. T.; MURTUGUDDE, R.; STRUTTON, P. G.; MANZANO-SARABIA, M.; WANG, H.; MITCHELL, B. G. Global correlations between winds and ocean chlorophyll. **Journal of Geophysical Research: Oceans**, Estados Unidos da América, v. 115, n. C12, p. 2010JC006500, dez. 2010.

KALTEH, A.; HJORTH, P.; BERNDTSSON, R. Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. **Environmental Modelling & Software**, Holanda, v. 23, n. 7, p. 835–845, jul. 2008.

KAM, J.; STOWERS, K.; KIM, S. Monitoring of Drought Awareness from Google Trends: A Case Study of the 2011–17 California Drought. **Weather, Climate, and Society**, Estados Unidos da América, v. 11, n. 2, p. 419–429, abr. 2019.

KARADIREK, I. E.; KARA, S.; YILMAZ, G.; MUHAMMETOGLU, A.; MUHAMMETOGLU, H. Implementation of Hydraulic Modelling for Water-Loss Reduction Through Pressure Management. **Water Resources Management**, Holanda, v. 26, n. 9, p. 2555–2568, jul. 2012.

KI-MOON, B.; GENERAL, U. S. **The human right to water and sanitation**. UN: 2010.

KIM, T.-W.; NAJJAR, R. G.; LEE, K. Influence of precipitation events on phytoplankton biomass in coastal waters of the eastern United States: Rain effect on phytoplankton biomass. **Global Biogeochemical Cycles**, Estados Unidos da América, v. 28, n. 1, p. 1–13, jan. 2014.

KOHONEN, T. Self-organized formation of topologically correct feature maps. **Biological Cybernetics**, Alemanha, v. 43, n. 1, p. 59–69, 1982.

KOHONEN, T. **MATLAB Implementation sand Applications of the Self-Organizing Map**. Helsinki: Unigrafia Oy, 2014.

KRAEMER, B. M.; ANNEVILLE, O.; CHANDRA, S.; DIX, M.; KUUSISTO, E.; LIVINGSTONE, D. M.; RIMMER, A.; SCHLADOW, S. G.; SILOW, E.; SITOKI, L. M.; TAMATAMAH, R.; VADEBONCOEUR, Y.; MCINTYRE, P. B. Morphometry and average temperature affect lake stratification responses to climate change. **Geophysical Research Letters**, Estados Unidos da América, v. 42, n. 12, p. 4981–4988, jun. 2015.

KRZYWINSKI, M.; ALTMAN, N. Points of Significance: Classification and regression trees. **Nature Methods**, Reino Unido, v. 14, p. 757–758, jul. 2017.

KUMANLIOGLU, A. A.; FISTIKOGLU, O. Performance Enhancement of a Conceptual Hydrological Model by Integrating Artificial Intelligence. **Journal of Hydrologic Engineering**, Estados Unidos da América, v. 24, n. 11, p. 04019047, nov. 2019.

KUMAR, P. Hydrocomplexity: Addressing water security and emergent environmental risks. **Water Resources Research**, Estados Unidos da América, v. 51, n. 7, p. 5827–5838, 2015.

LABADIE, J. W. Optimal Operation of Multireservoir Systems: State-of-the-Art Review. **Journal of Water Resources Planning and Management**, Estados Unidos da América, v. 130, n. 2, p. 93–111, mar. 2004.

LEE, D.; DERRIBLE, S. Predicting Residential Water Demand with Machine-Based Statistical Learning. **Journal of Water Resources Planning and Management**, Estados Unidos da América, v. 146, n. 1, p. 04019067, jan. 2020.

LI, T.; SUN, G.; YANG, C.; LIANG, K.; MA, S.; HUANG, L. Using self-organizing map for coastal water quality classification: Towards a better understanding of patterns and processes. **Science of The Total Environment**, Holanda, v. 628-629, p. 1446–1459, jul. 2018.

LI, T.; ZHANG, Y.; HE, B.; YANG, B.; HUANG, Q. Periodically hydrologic alterations decouple the relationships between physicochemical variables and chlorophyll-a in a dam-induced urban lake. **Journal of Environmental Sciences**, China, v. 99, p. 187–195, jan. 2021.

LIAW, A.; WIENER, M. Classification and Regression by randomForest. **R News**, v. 2, n. 3, p. 18–22, 2002. Disponível em: https://CRAN.R-project.org/doc/Rnews/.

LIMA NETO, I. E. Impact of artificial destratification on water availability of reservoirs in the Brazilian semiarid. **Anais da Academia Brasileira de Ciências**, Brasil, v. 91, n. 3, p. e20171022, 2019.

LIN, Y.; WANG, D.; WANG, G.; QIU, J.; LONG, K.; DU, Y.; XIE, H.; WEI, Z.; SHANGGUAN, W.; DAI, Y. A hybrid deep learning algorithm and its application to streamflow prediction. **Journal of Hydrology**, Holanda, v. 601, p. 126636, out. 2021.

LINDSAY, J.; DEAN, A. J.; SUPSKI, S. Responding to the Millennium drought: comparing domestic water cultures in three Australian cities. **Regional Environmental Change**, Holanda, v. 17, n. 2, p. 565–577, fev. 2017.

LINS, R.; MARTINEZ, J.-M.; MARQUES, D. M.; CIRILO, J.; FRAGOSO, C. Assessment of Chlorophyll-a Remote Sensing Algorithms in a Productive Tropical Estuarine-Lagoon System. **Remote Sensing**, Suíça, v. 9, n. 6, p. 516, maio 2017.

LIPPMANN, R. An introduction to computing with neural nets. **IEEE ASSP Magazine**, v. 4, n. 2, p. 4–22, 1987.

LIU, X.; FENG, J.; WANG, Y. Chlorophyll a predictability and relative importance of factors governing lake phytoplankton at different timescales. **Science of The Total Environment**, Holanda, v. 648, p. 472–480, jan. 2019.

LIU, Y.; ZHAO, J.; WANG, Z. Identifying determinants of urban water use using data mining approach. **Urban Water Journal**, Reino Unido, v. 12, n. 8, p. 618–630, nov. 2015.

LOPES, F. B.; BARBOSA, C. C. F.; NOVO, E. M. L. M.; ANDRADE, E. M.; CHAVES, L. C. G. Modelagem da qualidade das águas a partir de sensoriamento remoto hiperespectral. **Revista Brasileira de Engenharia Agrícola e Ambiental**, Brasil, v. 18, n. suppl, p. 13–19, 2014.

LOUCKS, D. P.; BEEK, E. v.; STEDINGER, J. R.; DIJKMAN, J. P. M.; VILLARS, M. T. **Water resource systems planning and management**: an introduction to methods, models, and applications. Cham, Suíça: Springer, 2017.

LUHMANN, N. **The reality of the mass media.** Stanford, California: Stanford University Press, 2000.

MA, X.; ZHANG, S.; MU, Q. How Do Residents Respond to Price under Increasing Block Tariffs? Evidence from Experiments in Urban Residential Water Demand in Beijing. **Water Resources Management**, Holanda, v. 28, n. 14, p. 4895–4909, nov. 2014.

MAIDMENT, D. R.; MIAOU, S.-P. Daily Water Use in Nine Cities. **Water Resources Research**, Estados Unidos da América, v. 22, n. 6, p. 845–851, jun. 1986.

MAMUN, M.; KIM, J.-J.; ALAM, M. A.; AN, K.-G. Prediction of Algal Chlorophyll-a and Water Clarity in Monsoon-Region Reservoir Using Machine Learning Approaches. **Water**, Suíça, v. 12, n. 1, p. 30, dez. 2019.

MANSUR, E. T.; OLMSTEAD, S. M. The value of scarce water: Measuring the inefficiency of municipal regulations. **Journal of Urban Economics**, Estados Unidos da América, v. 71, n. 3, p. 332–346, maio 2012.

MARTINEZ-ESPIñEIRA, R. Residential Water Demand in the Northwest of Spain. **Environmental and Resource Economics**, Holanda, v. 21, n. 2, p. 161–187, 2002.

MATIKINCA, P.; ZIERVOGEL, G.; ENQVIST, J. P. Drought response impacts on household water use practices in Cape Town, South Africa. **Water Policy**, Reino Unido, v. 22, n. 3, p. 483–500, jun. 2020.

MATOS, C.; TEIXEIRA, C. A.; BENTO, R.; VARAJãO, J.; BENTES, I. An exploratory study on the influence of socio-demographic characteristics on water end uses inside buildings. **Science of The Total Environment**, Holanda, v. 466-467, p. 467–474, jan. 2014.

MCDONALD, R. I.; WEBER, K.; PADOWSKI, J.; FLöRKE, M.; SCHNEIDER, C.; GREEN, P. A.; GLEESON, T.; ECKMAN, S.; LEHNER, B.; BALK, D.; BOUCHER, T.; GRILL, G.; MONTGOMERY, M. Water on an urban planet: Urbanization and the reach of urban water infrastructure. **Global Environmental Change**, Reino Unido, v. 27, p. 96–105, jul. 2014.

MEDEIROS, L. d. C.; MATTOS, A.; LüRLING, M.; BECKER, V. Is the future blue-green or brown? The effects of extreme events on phytoplankton dynamics in a semi-arid man-made lake. **Aquatic Ecology**, Holanda, v. 49, n. 3, p. 293–307, set. 2015.

MESQUITA, J. B. F.; LIMA NETO, I. E. L.; RAABE, A.; ARAÚJO, J. C. de. The influence of hydroclimatic conditions and water quality on evaporation rates of a tropical lake. **Journal of Hydrology**, Holanda, v. 590, p. 125456, nov. 2020.

MEYER, D.; DIMITRIADOU, E.; HORNIK, K.; WEINGESSEL, A.; LEISCH, F. **e1071**: Misc Functions of the Department of Statistics, Probability Theory Group, TU Wien. [*S.* l.], 2020. Disponível em: https://CRAN.R-project.org/package=e1071. Acesso em: 21 jan. 2023.

MILBORROW, S. **rpart.plot**: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'. ¢¢ [S. *l.*], 2020. Disponível em: https://CRAN.R-project.org/package=rpart.plot. Acesso em: 21 jan. 2023.

MILLY, P. C. D.; BETANCOURT, J.; FALKENMARK, M.; HIRSCH, R. M.; KUNDZEWICZ, Z. W.; LETTENMAIER, D. P.; STOUFFER, R. J. Stationarity Is Dead: Whither Water Management? **Science**, Estados Unidos da América, v. 319, n. 5863, p. 573–574, fev. 2008.

MOLINOS-SENANTE, M. Water rate to manage residential water demand with seasonality: peak-load pricing and increasing block rates approach. **Water Policy**, Reino Unido, v. 16, n. 5, p. 930–944, out. 2014.

MOLINOS-SENANTE, M.; DONOSO, G. Water scarcity and affordability in urban water pricing: A case study of Chile. **Utilities Policy**, Reino Unido, v. 43, p. 107–116, dez. 2016.

MOLNAR, C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Disponível em: https://christophm.github.io/interpretable-ml-book. Acesso em: 17 jan. 2022.

MONTEIRO, H.; ROSETA-PALMA, C. Pricing for scarcity? An efficiency analysis of increasing block tariffs: pricing for scarcity. **Water Resources Research**, Estados Unidos da América, v. 47, n. 6, jun. 2011.

MONTGOMERY, M. A.; ELIMELECH, M. Water And Sanitation in Developing Countries: Including Health in the Equation. **Environmental Science & Technology**, Estados Unidos da América, v. 41, n. 1, p. 17–24, jan. 2007.

MORIASI, D. N.; GITAU, M. W.; PAI, N.; DAGGUPATI, P. Hydrologic and Water Quality Models: Performance Measures and Evaluation Criteria. **Transactions of the ASABE**, v. 58, n. 6, p. 1763–1785, 2015.

MORTAZAVI-NAEINI, M.; KUCZERA, G.; CUI, L. Application of multiobjective optimization to scheduling capacity expansion of urban water resource systems. **Water Resources Research**, Estados Unidos da América, v. 50, n. 6, p. 4624–4642, jun. 2014.

MOURA, D. S.; LIMA NETO, I. E.; CLEMENTE, A.; OLIVEIRA, S.; PESTANA, C. J.; MELO, M. Aparecida de; CAPELO-NETO, J. Modeling phosphorus exchange between bottom sediment and water in tropical semiarid reservoirs. **Chemosphere**, Reino Unido, v. 246, p. 125686, maio 2020.

MSIZA, I. S.; NELWAMONDO, F. V.; MARWALA, T. Artificial Neural Networks and Support Vector Machines for water demand time series forecasting. *In:* **IEEE International Conference on Systems, Man and Cybernetics**. Montreal, QC, Canada: IEEE, 2007. p. 638–643.

MUSA, Z. N.; POPESCU, I.; MYNETT, A. A review of applications of satellite SAR, optical, altimetry and DEM data for surface water modelling, mapping and parameter estimation. **Hydrology and Earth System Sciences**, Alemanha, v. 19, n. 9, p. 3755–3769, set. 2015.

MUSOLESI, A.; NOSVELLI, M. Dynamics of residential water consumption in a panel of Italian municipalities. **Applied Economics Letters**, Reino Unido, v. 14, n. 6, p. 441–444, maio 2007.

MUñOZ-SABATER, J.; DUTRA, E.; AGUSTÍ-PANAREDA, A.; ALBERGEL, C.; ARDUINI, G.; BALSAMO, G.; BOUSSETTA, S.; CHOULGA, M.; HARRIGAN, S.; HERSBACH, H.; others. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. **Earth System Science Data**, Alemanha, v. 13, n. 9, p. 4349–4383, 2021.

NAWAZ, R.; REES, P.; CLARK, S.; MITCHELL, G.; MCDONALD, A.; KALAMANDEEN, M.; LAMBERT, C.; HENDERSON, R. Long-Term Projections of Domestic Water Demand: A Case Study of London and the Thames Valley. **Journal of Water Resources Planning and Management**, Estados Unidos da América, v. 145, n. 11, p. 05019017, nov. 2019.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. **Journal of the Royal Statistical Society. Series A (General)**, Reino Unido, v. 135, n. 3, p. 370, 1972.

NGUYEN, H.-Q.; HA, N.-T.; PHAM, T.-L. Inland harmful cyanobacterial bloom prediction in the eutrophic Tri An Reservoir using satellite band ratio and machine learning approaches. **Environmental Science and Pollution Research**, Alemanha, v. 27, n. 9, p. 9135–9151, mar. 2020.

OECD. **Mitigating Droughts and Floods in Agriculture: Policy Lessons and Approaches**. Paris: [*S. n.*], 2016. Disponível em: https://www.oecd-ilibrary.org/agriculture-and-food/mitigating-droughts-and-floods-in-agriculture_9789264246744-en. Acesso em: 21 nov. 2021.

OLIVEIRA, G.; SCAZUFCA, P.; SAYON, P. L. **Desafios Para a Disponibilidade Hídrica e ao Avanço da Eficiência do Saneamento Básico**. São Paulo, 2022. 67 p.

OLMSTEAD, S. M. Climate change adaptation and water resource management: A review of the literature. **Energy Economics**, Holanda, v. 46, p. 500–509, nov. 2014.

OLMSTEAD, S. M.; HANEMANN, W. M.; STAVINS, R. N. Water demand under alternative price structures. **Journal of Environmental Economics and Management**, Estados Unidos da América, v. 54, n. 2, p. 181–198, set. 2007.

OLMSTEAD, S. M.; STAVINS, R. N. Comparing price and nonprice approaches to urban water conservation: approaches to urban water conservation. **Water Resources Research**, Estados Unidos da América, v. 45, n. 4, abr. 2009.

OLSSON, J. A.; ANDERSSON, L. Possibilities and problems with the use of models as a communication tool in water resource management. *In:* CRASWELL, E.; BONNELL, M.; BOSSIO, D.; DEMUTH, S.; GIESEN, N. V. D. (Ed.). **Integrated Assessment of Water Resources and Global Change: A North-South Analysis**. Dordrecht: Springer Netherlands, 2007. p. 97–110. Disponível em: https://doi.org/10.1007/978-1-4020-5591-1_7. Acesso em: 20

ORLOWSKY, B.; SENEVIRATNE, S. I. Global changes in extreme events: regional and seasonal dimension. **Climatic Change**, Holanda, v. 110, n. 3-4, p. 669–696, fev. 2012.

OUYANG, Y.; WENTZ, E. A.; RUDDELL, B. L.; HARLAN, S. L. A Multi-Scale Analysis of Single-Family Residential Water Use in the Phoenix Metropolitan Area. **Journal of the American Water Resources Association**, Estados Unidos da América, v. 50, n. 2, p. 448–467, abr. 2014.

PACHECO, C. H. A.; LIMA NETO, I. E. Effect of Artificial Circulation on the Removal Kinetics of Cyanobacteria in a Hypereutrophic Shallow Lake. **Journal of Environmental Engineering**, Japão, v. 143, n. 12, p. 06017010, dez. 2017.

PADULANO, R.; GIUDICE, G. D. A Mixed Strategy Based on Self-Organizing Map for Water Demand Pattern Profiling of Large-Size Smart Water Grid Data. **Water Resources Management**, Holanda, v. 32, n. 11, p. 3671–3685, set. 2018.

PAERL, H. W.; OTTEN, T. G. Harmful Cyanobacterial Blooms: Causes, Consequences, and Controls. **Microbial Ecology**, Estados Unidos da América, v. 65, n. 4, p. 995–1010, maio 2013.

PAL, I.; ANDERSON, B. T.; SALVUCCI, G. D.; GIANOTTI, D. J. Shifting seasonality and increasing frequency of precipitation in wet and dry seasons across the U.S.: U.S. precipitation seasonality change. **Geophysical Research Letters**, Estados Unidos da América, v. 40, n. 15, p. 4030–4035, ago. 2013.

PAPACHARALAMPOUS, G. A.; TYRALIS, H. Evaluation of random forests and Prophet for daily streamflow forecasting. **Advances in Geosciences**, Alemanha, v. 45, p. 201–208, ago. 2018.

PARANDVASH, G. H.; CHANG, H. Analysis of long-term climate change on per capita water demand in urban versus suburban areas in the Portland metropolitan area, USA. **Journal of Hydrology**, Estados Unidos da América, v. 538, p. 574–586, jul. 2016.

PEREIRA, M. V. F.; PINTO, L. M. V. G. Multi-stage stochastic optimization applied to energy planning. **Mathematical Programming**, Alemanha, v. 52, n. 1-3, p. 359–375, maio 1991.

PERRY, C.; STEDUTO, P.; KARAJEH, F. **Does improved irrigation technology save water?** A review of the evidence. Cairo, Food and Agriculture Organization of the United Nations, 2017, 42.

PESIC, R.; JOVANOVIC, M.; JOVANOVIC, J. Seasonal water pricing using meteorological data: case study of Belgrade. **Journal of Cleaner Production**, Reino Unido, v. 37, p. 299–303, dez. 2012.

PNUD; IPEA; FJP. **Atlas do desenvolvimento humano nas regiões metropolitanas**. [*S. l.*], 2014. Disponível em: http://atlasbrasil.org.br/2013/data/rawData/publicacao_atlas_rm_en.pdf. Acesso em: 14 mar. 2020.

PNUD, B. **Programa das Nações Unidas para o desenvolvimento**. Rio de Janeiro, 2012. Disponível em: http://www.atlasbrasil.org.br. Acesso em: 14 mar. 2020.

POLEBITSKI, A. S.; PALMER, R. N. Seasonal Residential Water Demand Forecasting for Census Tracts. **Journal of Water Resources Planning and Management**, Estados Unidos da América, v. 136, n. 1, p. 27–36, jan. 2010.

PONTES FILHO, J. D.; SOUZA FILHO, F. A. S.; MARTINS, E. S. P. R.; STUDART, T. M. C. Copula-Based Multivariate Frequency Analysis of the 2012–2018 Drought in Northeast Brazil. Water, Suíça, v. 12, n. 3, p. 834, mar. 2020.

PORTO, V. C.; SOUZA FILHO, F. A.; CARVALHO, T. M. N.; STUDART, T. M. d. C.; PORTELA, M. M. A GLM copula approach for multisite annual streamflow generation. **Journal of Hydrology**, Estados Unidos da América, v. 598, p. 126226, jul. 2021.

PRIGOGINE, I.; STENGERS, I. **A Nova Aliança**. [*S. l.*]: Universidade de Brasília, 1991.

PULIDO-CALVO, I.; MONTESINOS, P.; ROLDáN, J.; RUIZ-NAVARRO, F. Linear regressions and neural approaches to water demand forecasting in irrigation districts with telemetry systems. **Biosystems Engineering**, Estados Unidos da América, v. 97, n. 2, p. 283–293, jun. 2007.

QI, J.; LIU, H.; LIU, X.; ZHANG, Y. Spatiotemporal evolution analysis of time-series land use change using self-organizing map to examine the zoning and scale effects. **Computers, Environment and Urban Systems**, Reino Unido, v. 76, p. 11–23, jul. 2019.

QIN, B.; ZHOU, J.; ELSER, J. J.; GARDNER, W. S.; DENG, J.; BROOKES, J. D. Water Depth Underpins the Relative Roles and Fates of Nitrogen and Phosphorus in Lakes. **Environmental Science & Technology**, Estados Unidos da América, v. 54, n. 6, p. 3191–3198, mar. 2020.

QUESNEL, K. J.; AJAMI, N. K. Changes in water consumption linked to heavy news media coverage of extreme climatic events. **Science Advances**, Estados Unidos da América, v. 3, n. 10, p. e1700784, out. 2017.

RASIFAGHIHI, N.; LI, S.; HAGHIGHAT, F. Forecast of urban water consumption under the impact of climate change. **Sustainable Cities and Society**, Holanda, v. 52, p. 101848, jan. 2020.

RAULINO, J. B. S.; SILVEIRA, C. S.; LIMA NETO, I. E. Assessment of climate change impacts on hydrology and water quality of large semi-arid reservoirs in Brazil. **Hydrological Sciences Journal**, Reino Unido, v. 66, n. 8, p. 1321–1336, jun. 2021.

REED, R.; MARKS, R. J. **Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks**. The MIT Press, 1999. ISBN 978-0-262-28221-5. Disponível em: https://direct.mit.edu/books/book/2736/neural-smithingsupervised-learning-in-feedforward.

REICHWALDT, E. S.; GHADOUANI, A. Effects of rainfall patterns on toxic cyanobacterial blooms in a changing climate: Between simplistic scenarios and complex dynamics. **Water Research**, Reino Unido, v. 46, n. 5, p. 1372–1393, abr. 2012.

RIBEIRO, F. W.; SILVA, S. M. O.; SOUZA FILHO, F. A.; CARVALHO, T. M. N.; LOPES, T. M. X. M. Diversification of urban water supply: An assessment of social costs and water production costs. **Water Policy**, Reino Unido, v. 24, n. 6, p. 980–997, jun. 2022.

RIETVELD, P.; ROUWENDAL, J.; ZWART, B. Block Rate Pricing of Water in Indonesia: An Analysis of Welfare Effects. **Bulletin of Indonesian Economic Studies**, Reino Unido, v. 36, n. 3, p. 73–92, dez. 2000.

RINAUDO, J.-D.; NEVERRE, N.; MONTGINOUL, M. Simulating the Impact of Pricing Policies on Residential Water Demand: A Southern France Case Study. **Water Resources Management**, Holanda, v. 26, n. 7, p. 2057–2068, maio 2012.

ROCHA, M. J. D.; LIMA NETO, I. E. Modeling flow-related phosphorus inputs to tropical semiarid reservoirs. **Journal of Environmental Management**, Estados Unidos da América, v. 295, p. 113123, out. 2021.

ROCHA, M. J. D.; LIMA NETO, I. E. Phosphorus mass balance and input load estimation from the wet and dry periods in tropical semiarid reservoirs. **Environmental Science and Pollution Research**, Alemanha, p. 1–20, set. 2021.

ROCHA, M. J. D.; LIMA NETO, I. E. Internal phosphorus loading and its driving factors in the dry period of Brazilian semiarid reservoirs. **Journal of Environmental Management**, Estados Unidos da América, v. 312, p. 114983, jun. 2022.

ROCHA, M. J. D.; LIMA NETO, I. E. Phosphorus mass balance and input load estimation from the wet and dry periods in tropical semiarid reservoirs. **Environmental Science and Pollution Research**, Alemanha, v. 29, n. 7, p. 10027–10046, fev. 2022.

ROCHA, S. M. G.; MESQUITA, J. B. F.; LIMA NETO, I. E. Análise da modelagem das relações entre nutrientes e fitoplâncton em reservatórios do Ceará. **Revista Brasileira de Ciências Ambientais (Online)**, Brasil, n. 54, p. 134–147, mar. 2020.

ROMANO, G.; SALVATI, N.; GUERRINI, A. Estimating the Determinants of Residential Water Demand in Italy. **Water**, Suíça, v. 6, n. 10, p. 2929–2945, set. 2014.

ROSA, J. G. **Grande sertão: veredas**. Nova edição. São Paulo (S.P.): Companhia das Letras, 2019. ISBN 978-85-359-3198-3.

ROSS, M. R. V.; TOPP, S. N.; APPLING, A. P.; YANG, X.; KUHN, C.; BUTMAN, D.; SIMARD, M.; PAVELSKY, T. M. AquaSat: A Data Set to Enable Remote Sensing of Water Quality for Inland Waters. **Water Resources Research**, Estados Unidos da América, v. 55, n. 11, p. 10012–10025, nov. 2019.

ROUGÉ, C.; TILMANT, A. Using stochastic dual dynamic programming in problems with multiple near-optimal solutions. **Water Resources Research**, Estados Unidos da América, v. 52, n. 5, p. 4151–4163, maio 2016.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, Holanda, v. 20, p. 53–65, nov. 1987.

RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. **Nature Machine Intelligence**, Suíça, v. 1, n. 5, p. 206–215, maio 2019.

RUIJS, A. Welfare and Distribution Effects of Water Pricing Policies. **Environmental and Resource Economics**, Holanda, v. 43, n. 2, p. 161–182, jun. 2009.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **Nature**, Reino Unido, v. 323, n. 6088, p. 533–536, out. 1986.

SAKAMOTO, M. Primary production by phytoplankton community in some Japanese lakes and its dependence on lake depth. **Archiv Fur Hydrobiologie**, v. 62, p. 1–28, 1966.

SANT'ANA, D.; MAZZEGA, P. Socioeconomic analysis of domestic water end-use consumption in the Federal District, Brazil. **Sustainable Water Resources Management**, Suíça, v. 4, n. 4, p. 921–936, dez. 2018.

SAURI, D. The decline of water consumption in Spanish cities: structural and contingent factors. **International Journal of Water Resources Development**, Reino Unido, v. 36, n. 6, p. 909–925, nov. 2020.

SCHLEICH, J.; HILLENBRAND, T. Determinants of residential water demand in Germany. **Ecological Economics**, Holanda, v. 68, n. 6, p. 1756–1769, abr. 2009.

SCHÄFER, B.; BECK, C.; RHYS, H.; SOTERIOU, H.; JENNINGS, P.; BEECHEY, A.; HEPPELL, C. M. Machine learning approach towards explaining water quality dynamics in an urbanised river. **Scientific Reports**, Reino Unido, v. 12, n. 1, p. 12346, jul. 2022.

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**, Reino Unido, v. 37, n. 1, p. 112–128, fev. 2010.

SHARMA, S. K.; VAIRAVAMOORTHY, K. Urban water demand management: prospects and challenges for the developing countries. **Water and Environment Journal**, Estados Unidos da América, v. 23, n. 3, p. 210–218, set. 2009.

SHEN, C. A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. **Water Resources Research**, Estados Unidos da América, v. 54, n. 11, p. 8558–8593, 2018.

SHEN, J.; QIN, Q.; WANG, Y.; SISSON, M. A data-driven modeling approach for simulating algal blooms in the tidal freshwater of James River in response to riverine nutrient loading. **Ecological Modelling**, Holanda, v. 398, p. 44–54, abr. 2019.

SIT, M.; DEMIRAY, B. Z.; XIANG, Z.; EWING, G. J.; SERMET, Y.; DEMIR, I. A comprehensive review of deep learning applications in hydrology and water resources. **Water Science and Technology**, Reino Unido, v. 82, n. 12, p. 2635–2670, ago. 2020.

SOLOMATINE, D.; SEE, L.; ABRAHART, R. Data-Driven Modelling: Concepts, Approaches and Experiences. In: ABRAHART, R. J.; SEE, L. M.; SOLOMATINE, D. P. (Ed.). **Practical Hydroinformatics**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. v. 68, p. 17–30.

SOUZA FILHO, F. A.; STUDART, T. M. C.; PONTES FILHO, J. D.; MARTINS, E. S. P. R.; AYRIMORAES, S. R.; PESSOA, C. A. P.; ROLIM, L. Z. R.; ARAUJO JUNIOR, L. M.; SILVA, S. M. O.; CARVALHO, T. M. N.; AQUINO, S. H. S.. Integrated proactive drought management in hydrosystems and cities: building a nine-step participatory planning methodology. Natural Hazards, Holanda, v. 115, n. 3, p. 2179–2204, fev. 2023.

SPERLING, M. v. **Introdução à Qualidade das Águas e ao Tratamento de Esgotos**. Belo Horizonte: DESA-UFMG, 2005.

STEFANIDIS, K.; VARLAS, G.; VOURKA, A.; PAPADOPOULOS, A.; DIMITRIOU, E. Delineating the relative contribution of climate related variables to chlorophyll-a and phytoplankton biomass in lakes using the ERA5-Land climate reanalysis data. **Water Research**, Reino Unido, v. 196, p. 117053, maio 2021.

STOCKWELL, J. D.; DOUBEK, J. P.; ADRIAN, R.; ANNEVILLE, O.; CAREY, C. C.; CARVALHO, L.; DOMIS, L. N. D. S.; DUR, G.; FRASSL, M. A.; GROSSART, H.; IBELINGS, B. W.; LAJEUNESSE, M. J.; LEWANDOWSKA, A. M.; LLAMES, M. E.; MATSUZAKI, S. S.; NODINE, E. R.; NõGES, P.; PATIL, V. P.; POMATI, F.; RINKE, K.; RUDSTAM, L. G.; RUSAK, J. A.; SALMASO, N.; SELTMANN, C. T.; STRAILE, D.; THACKERAY, S. J.; THIERY, W.; URRUTIA-CORDERO, P.; VENAIL, P.; VERBURG, P.; WOOLWAY, R. I.; ZOHARY, T.;

ANDERSEN, M. R.; BHATTACHARYA, R.; HEJZLAR, J.; JANATIAN, N.; KPODONU, A. T. N. K.; WILLIAMSON, T. J.; WILSON, H. L. Storm impacts on phytoplankton community dynamics in lakes. **Global Change Biology**, Reino Unido, v. 26, n. 5, p. 2756–2784, maio 2020.

STROBL, C.; BOULESTEIX, A.-L.; ZEILEIS, A.; HOTHORN, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. **BMC Bioinformatics**, Reino Unido, v. 8, n. 1, p. 25, dez. 2007.

SU, S.; XIAO, R.; XU, X.; ZHANG, Z.; MI, X.; WU, J. Multi-scale spatial determinants of dissolved oxygen and nutrients in Qiantang River, China. **Regional Environmental Change**, Alemanha, v. 13, n. 1, p. 77–89, maio 2012.

TAO, L.; HE, X.; LI, J.; YANG, D. A multiscale long short-term memory model with attention mechanism for improving monthly precipitation prediction. **Journal of Hydrology**, Estados Unidos da América, v. 602, p. 126815, nov. 2021.

THERNEAU, T.; ATKINSON, B. **rpart**: Recursive Partitioning and Regression Trees. [S. l.], 2019. Disponível em: https://CRAN.R-project.org/package=rpart. Acesso em: 8 jan. 2022.

TONG, Y.; XU, X.; ZHANG, S.; SHI, L.; ZHANG, X.; WANG, M.; QI, M.; CHEN, C.; WEN, Y.; ZHAO, Y.; ZHANG, W.; LU, X. Establishment of season-specific nutrient thresholds and analyses of the effects of nutrient management in eutrophic lakes through statistical machine learning. **Journal of Hydrology**, Estados Unidos da América, v. 578, p. 124079, nov. 2019.

TORRES, M. E.; COLOMINAS, M. A.; SCHLOTTHAUER, G.; FLANDRIN, P. A complete ensemble empirical mode decomposition with adaptive noise. *In:* International Conference on Acoustics, Speech and Signal Processing. 11., 2011, Prague. **Proceedings [...]**. Prague, Czech Republic: IEEE, 2011. p. 4144-4147.

TRINDADE, B.; REED, P.; CHARACKLIS, G. Deeply uncertain pathways: Integrated multi-city regional water supply infrastructure investment and portfolio management. **Advances in Water Resources**, Reino Unido, v. 134, p. 103442, dez. 2019.

TYRALIS, H.; PAPACHARALAMPOUS, G.; LANGOUSIS, A. A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources. **Water**, Suíça, v. 11, n. 5, p. 910, abr. 2019.

UNESCO. **Nature-based solutions for water**: Development Report. [*S. l.*], 2018.

UNESCO. **The United Nations World Development Report**: Leaving no one behind. [*S. l.*], 2019.

URRESTARAZU, L. P.; BURT, C. M. Characterization of Pumps for Irrigation in Central California: Potential Energy Savings. **Journal of Irrigation and Drainage**

Engineering, Reino Unido, v. 138, n. 9, p. 815–822, set. 2012.

VELDKAMP, T.; WADA, Y.; AERTS, J.; DöLL, P.; GOSLING, S. N.; LIU, J.; MASAKI, Y.; OKI, T.; OSTBERG, S.; POKHREL, Y.; SATOH, Y.; KIM, H.; WARD, P. J. Water scarcity hotspots travel downstream due to human interventions in the 20th and 21st century. **Nature Communications**, Reino Unido, v. 8, n. 1, p. 15697, jun. 2017.

VESANTO, J.; ALHONIEMI, E. Clustering of the self-organizing map. **IEEE Transactions on Neural Networks**, Estados Unidos da América, v. 11, n. 3, p. 586–600, maio 2000.

VIJAI, P.; SIVAKUMAR, P. B. Performance comparison of techniques for water demand forecasting. **Procedia Computer Science**, Holanda, v. 143, p. 258–266, 2018.

VILLARIN, M. C.; RODRIGUEZ-GALIANO, V. F. Machine Learning for Modeling Water Demand. **Journal of Water Resources Planning and Management**, Estados Unidos da América, v. 145, n. 5, p. 04019017, maio 2019.

VOSEN, S.; SCHMIDT, T. Forecasting private consumption: survey-based indicators vs. Google trends. **Journal of Forecasting**, Reino Unido, v. 30, n. 6, p. 565–578, set. 2011.

VÖRÖSMARTY, C. J.; MCINTYRE, P. B.; GESSNER, M. O.; DUDGEON, D.; PRUSEVICH, A.; GREEN, P.; GLIDDEN, S.; BUNN, S. E.; SULLIVAN, C. A.; LIERMANN, C. R.; DAVIES, P. M. Global threats to human water security and river biodiversity. **Nature**, Reino Unido, v. 467, n. 7315, p. 555–561, set. 2010.

WANG, X.; YANG, W. Water quality monitoring and evaluation using remote sensing techniques in China: a systematic review. **Ecosystem Health and Sustainability**, Reino Unido, v. 5, n. 1, p. 47–56, jan. 2019.

WARD, F. A.; PULIDO-VELAZQUEZ, M. Water conservation in irrigation can increase water use. **Proceedings of the National Academy of Sciences**, v. 105, n. 47, p. 18215–18220, nov. 2008.

WARD, F. A.; PULIDO-VELáZQUEZ, M. Efficiency, equity, and sustainability in a water quantity–quality optimization model in the Rio Grande basin. **Ecological Economics**, Holanda, v. 66, n. 1, p. 23–37, maio 2008.

WEESER, B.; KROESE, J. S.; JACOBS, S.; NJUE, N.; KEMBOI, Z.; RAN, A.; RUFINO, M.; BREUER, L. Citizen science pioneers in Kenya – A crowdsourced approach for hydrological monitoring. **Science of The Total Environment**, Reino Unido, v. 631-632, p. 1590–1599, ago. 2018.

WEHRENS, R.; KRUISSELBRINK, J. Flexible Self-Organizing Maps in kohonen 3.0. **Journal of Statistical Software,** v. 87, n. 7, p. 1–18, 2018.

WEI, C.-C.; HSU, N.-S. Derived operating rules for a reservoir operation system: Comparison of decision trees, neural decision trees and fuzzy decision trees. **Water Resources Research**, Estados Unidos da América, v. 44, n. 2, 2008.

WEI, T.; SIMKO, V. **corrplot**: Visualization of a Correlation Matrix. 2017. Disponível em: https://github.com/taiyun/corrplot. Acesso em: 27 fev. 2021.

WEI, W.; YAN, Z.; TONG, X.; HAN, Z.; MA, M.; YU, S.; XIA, J. Seasonal prediction of summer extreme precipitation over the Yangtze River based on random forest. **Weather and Climate Extremes**, Holanda, v. 37, p. 100477, set. 2022.

WHEELER, S. A.; ZUO, A.; LOCH, A. Watering the farm: Comparing organic and conventional irrigation water use in the Murray–Darling Basin, Australia. **Ecological Economics**, Holanda, v. 112, p. 78–85, abr. 2015.

WHITTINGTON, D.; NAUGES, C. An Assessment of the Widespread Use of Increasing Block Tariffs in the Municipal Water Supply Sector. *In:* Oxford Research Encyclopedia of Global Public Health. Oxford University Press, Oxford, 2020. p. 1-20. Disponível em: https://oxfordre.com/publichealth/view/10.1093/acrefore/9780190632366.001.0001/acrefore-9780190632366-e-243. Acesso em: 27 jan. 2021.

WICKHAM, H. **ggplot2**: Elegant Graphics for Data Analysis. Springer-Verlag New York, [*S.* l.], 2016. Disponível em: https://ggplot2.tidyverse.org. Acesso em: 23 out. 2021.

WICKHAM, H. **tidyr**: Tidy Messy Data. [*S.* l.], 2020. Disponível em: https://CRAN.R-project.org/package=tidyr. Acesso em: 27 jan. 2021.

WICKHAM, H.; FRANÇOIS, R.; HENRY, L.; MüLLER, K. dplyr: A Grammar of Data Manipulation. [*S.* l.], 2022.

WIEGAND, M. C.; NASCIMENTO, A. T. P. d.; COSTA, A. C.; LIMA, I. E. Avaliação de nutriente limitante da produção algal em reservatórios do semiárido brasileiro. **Revista Brasileira de Ciências Ambientais (Online)**, Brasil, v. 55, n. 4, p. 456–478, 2020.

WIEGAND, M. C.; NASCIMENTO, A. T. P.; COSTA, A. C.; LIMA NETO, I. E. Trophic state changes of semi-arid reservoirs as a function of the hydro-climatic variability. **Journal of Arid Environments**, Estados Unidos da América, v. 184, p. 104321, jan. 2021.

WITTEN, I. H.; FRANK, E.; HALL, M. A.: Practical Machine Learning Tools and Techniques. Elsevier, 2011. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/C20090197155. Acesso em: 8 jan. 2021.

WU, Z.; HUANG, N. E. Ensemble Empirical Mode Decomposition: A Noise-Assisted Data Analysis Method. **Advances in Adaptive Data Analysis**, Singapura, v. 01, n. 01, p. 1–41, jan. 2009.

XIAO, Y.; FANG, L.; HIPEL, K. W. Agent-Based Modeling Approach to Investigating the Impact of Water Demand Management. **Journal of Water Resources Planning and Management**, Estados Unidos da América, v. 144, n. 3, p. 04018006, mar. 2018.

XIONG, W.; LI, Y.; ZHANG, W.; YE, Q.; ZHANG, S.; HOU, X. Integrated multi-objective optimization framework for urban water supply systems under alternative climates and future policy. **Journal of Cleaner Production**, Reino Unido, v. 195, p. 640–650, set. 2018.

XU, W.; CHEN, J.; ZHANG, X. J.; XIONG, L.; CHEN, H. A framework of integrating heterogeneous data sources for monthly streamflow prediction using a state-of-the-art deep learning model. **Journal of Hydrology**, Holanda, v. 614, p. 128599, nov. 2022.

YAJIMA, H.; DEROT, J. Application of the Random Forest model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases. **Journal of Hydroinformatics**, Reino Unido, v. 20, n. 1, p. 206–220, jan. 2018.

YANG, T.; GAO, X.; SOROOSHIAN, S.; LI, X. Simulating California reservoir operation using the classification and regression-tree algorithm combined with a shuffled cross-validation scheme. **Water Resources Research**, Estados Unidos da América, v. 52, n. 3, p. 1626–1651, 2016.

YANG, X.-e.; WU, X.; HAO, H.-l.; HE, Z.-l. Mechanisms and assessment of water eutrophication. **Journal of Zhejiang University SCIENCE B**, China, v. 9, n. 3, p. 197–209, mar. 2008.

YEH, J.-R.; SHIEH, J.-S.; HUANG, N. E. Complementary Ensemble Empirical Mode Decomposition: A Novel Noise Enhanced Data Analysis Method. **Advances in Adaptive Data Analysis**, Singapura, v. 02, n. 02, p. 135–156, abr. 2010.

YOUSEFI, P.; SHABANI, S.; MOHAMMADI, H.; NASER, G. Gene Expression Programing in Long Term Water Demand Forecasts Using Wavelet Decomposition. **Procedia Engineering**, Holanda, v. 186, p. 544–550, 2017.

YU, X.; SHEN, J.; DU, J. A Machine-Learning-Based Model for Water Quality in Coastal Waters, Taking Dissolved Oxygen and Hypoxia in Chesapeake Bay as an Example. **Water Resources Research**, Estados Unidos da América, v. 56, n. 9, p. e2020WR027227, set. 2020.

ZHANG, B.; FANG, K. H.; BAERENKLAU, K. A. Have Chinese water pricing reforms reduced urban residential water demand? **Water Resources Research**, Estados Unidos da América, v. 53, n. 6, p. 5057–5069, jun. 2017.

ZHAO, Y.; HAN, Q.; DING, C.; HUANG, Y.; LIAO, J.; CHEN, T.; FENG, S.; ZHOU, L.; ZHANG, Z.; CHEN, Y.; YUAN, S.; YUAN, M. Effect of Low Temperature on Chlorophyll Biosynthesis and Chloroplast Biogenesis of Rice Seedlings during Greening. **International Journal of Molecular Sciences**, Suíça, v. 21, n. 4, p. 1390, fev. 2020.

ZHUANG, W. Eco-environmental impact of inter-basin water transfer projects: a review. **Environmental Science and Pollution Research**, Alemanha, v. 23, n. 13, p. 12867–12879, jul. 2016.

ZIEGLER, A.; KÖNIG, I. R. Mining data with random forests: current options for real-world applications: Mining data with random forests. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Estados Unidos da América, v. 4, n. 1, p. 55–63, jan. 2014.

ZUBAIDI, S. L.; ORTEGA-MARTORELL, S.; KOT, P.; ALKHADDAR, R. M.; ABDELLATIF, M.; GHARGHAN, S. K.; AHMED, M. S.; HASHIM, K. A Method for Predicting Long-Term Municipal Water Demands Under Climate Change. **Water Resources Management**, Holanda, v. 34, n. 3, p. 1265–1279, fev. 2020.

ZUO, G.; LUO, J.; WANG, N.; LIAN, Y.; HE, X. Decomposition ensemble model based on variational mode decomposition and long short-term memory for streamflow forecasting. **Journal of Hydrology**, Estados Unidos da América, v. 585, p. 124776, jun. 2020.

APPENDIX A – PUBLICATIONS

A.1 Publications in scientific journals

A.1.1 Included in the PhD thesis

1. NUNES CARVALHO, T. M.; SOUZA FILHO, F. DE A.; COSTA PORTO, V. Urban Water Demand Modeling Using Machine Learning Techniques: Case Study of Fortaleza, Brazil. Journal of Water Resources Planning and Management, v. 147, 1 jan. 2021.
2. CARVALHO, T. M. N.; DE ASSIS DE SOUZA FILHO, F. Variational Mode Decomposition Hybridized With Gradient Boost Regression for Seasonal Forecast of Residential Water Demand. Water Resources Management, v. 35, n. 10, p. 3431–3445, ago. 2021.
3. NUNES CARVALHO, T. M.; DE SOUZA FILHO, F. DE A. A data-driven model to evaluate the medium-term effect of contingent pricing policies on residential water demand. Environmental Challenges, v. 3, p. 100033, 1 abr. 2021.
4. NUNES CARVALHO, T. M.; LIMA NETO, I. E.; SOUZA FILHO, F. DE A. Uncovering the influence of hydrological and climate variables in chlorophyll-A concentration in tropical reservoirs with machine learning. Environmental Science and Pollution Research, v. 29, n. 49, p. 74967–74982, out. 2022.

A.1.2 Not included in the PhD thesis

5. NUNES CARVALHO, T. M.; DE SOUZA FILHO, F. D. A.; MEDEIROS DE SABÓIA, M. A. Performance of rainwater tanks for runoff reduction under climate change scenarios: a case study in Brazil. Urban Water Journal, v. 17, n. 10, p. 912–922, 25 nov. 2020.
6. CARVALHO, T. M. et al. Índice de vulnerabilidade à COVID-19: uma aplicação para a cidade de Fortaleza (CE), Brasil. Engenharia Sanitaria e Ambiental, v. 26, n. 4, p. 731–739, ago. 2021.

A.1.3 Co-authored publications

7. XAVIER, L. C. P. et al. Use of Machine Learning in Evaluation of Drought Perception in Irrigated Agriculture: The Case of an Irrigated Perimeter in Brazil. Water, v. 12, n. 6, p. 1546, 28 maio 2020.
8. MARQUES DE OLIVEIRA, L. et al. Forecasting Urban Water Demand Using Cellular

Automata. Water, v. 12, n. 7, p. 2038, 17 jul. 2020.

9. PORTO, V. C. et al. A GLM copula approach for multisite annual streamflow generation. Journal of Hydrology, v. 598, p. 126226, 1 jul. 2021.
10. RIBEIRO, F. W. et al. Diversification of urban water supply: An assessment of social costs and water production costs. Water Policy, v. 24, n. 6, p. 980–997, 1 jun. 2022.
11. DE ASSIS SOUZA FILHO, F. et al. Integrated proactive drought management in hydrosystems and cities: building a nine-step participatory planning methodology. Natural Hazards, 17 nov. 2022.
12. CARNEIRO, B. L. D. S. et al. Hydrological risk of dam failure under climate change. RBRH, v. 27, p. e19, 2022.

A.2 Book chapters

13. CARVALHO, TAÍS MARIA NUNES; SOUZA FILHO, F. A. . Previsão Sazonal de Demanda Residencial de Água de Fortaleza. SECAS E CHEIAS: Modelagem e Adaptação aos extremos hidrológicos no contexto da variabilidade e mudança do clima. 1ed., 2022. Previsão Sazonal de Demanda Residencial de Água de Fortaleza
14. CARVALHO, TAÍS MARIA NUNES; SOUZA FILHO, F. A. . SECAS E CHEIAS: Modelagem e Adaptação aos extremos hidrológicos no contexto da variabilidade e mudança do clima. SECAS E CHEIAS: Modelagem e Adaptação aos extremos hidrológicos no contexto da variabilidade e mudança do clima. 1ed.: , 2022.

A.3 Conference papers

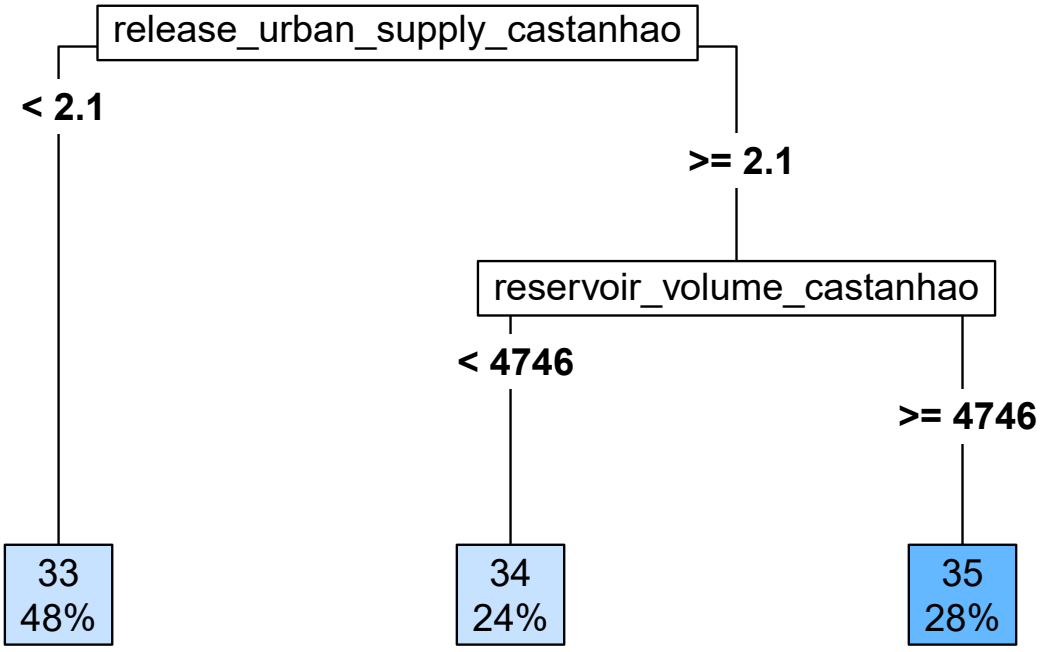
15. CARVALHO, T. M. N. et al. Integrated model of capacity expansion and operation of water supply systems including non-conventional water sources). Spain: European Water Resources Association EWRA, 2019.
16. CARVALHO, Taís Maria Nunes; SOUZA FILHO, Francisco de Assis de; LOPES, Tereza Margarida Xavier de Melo. Detecção de secas e visualização de padrões climáticos com aprendizado de máquina. In: SIMPÓSIO BRASILEIRO DE RECURSOS HÍDRICOS, XXIV., 21 a 26 nov. 2021, Belo Horizonte-MG. Anais[...], Belo Horizonte-MG., 2021.

A.4 Technical reports

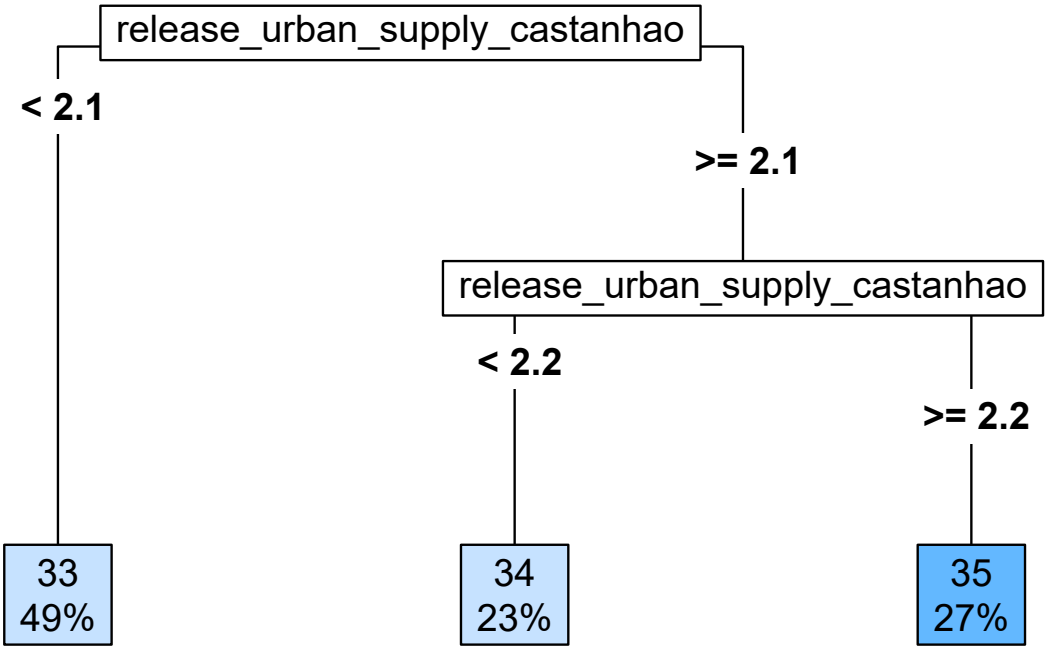
17. NUNES CARVALHO, T. M.; SOUZA FILHO, F. A.; OLIVEIRA DA SILVA, S. M. Modelo de Otimização da Expansão da Capacidade do Sistema Integrado de Abastecimento de Água de Fortaleza. Fortaleza: UFC, 2022.
18. NUNES CARVALHO, T. M.; SOUZA FILHO, F. A.; OLIVEIRA DA SILVA, S. M. Simulação do Sistema Integrado de Abastecimento de Água de Fortaleza. Fortaleza: UFC, 2022.
19. NUNES CARVALHO, T. M.; SOUZA FILHO, F. A.; OLIVEIRA DA SILVA, S. M. Manual do Modelo de Otimização da Expansão da Capacidade do Sistema Integrado de Abastecimento de Água de Fortaleza. Fortaleza: UFC, 2022.

APPENDIX B - DECISION TREES OBTAINED IN CHAPTER 7

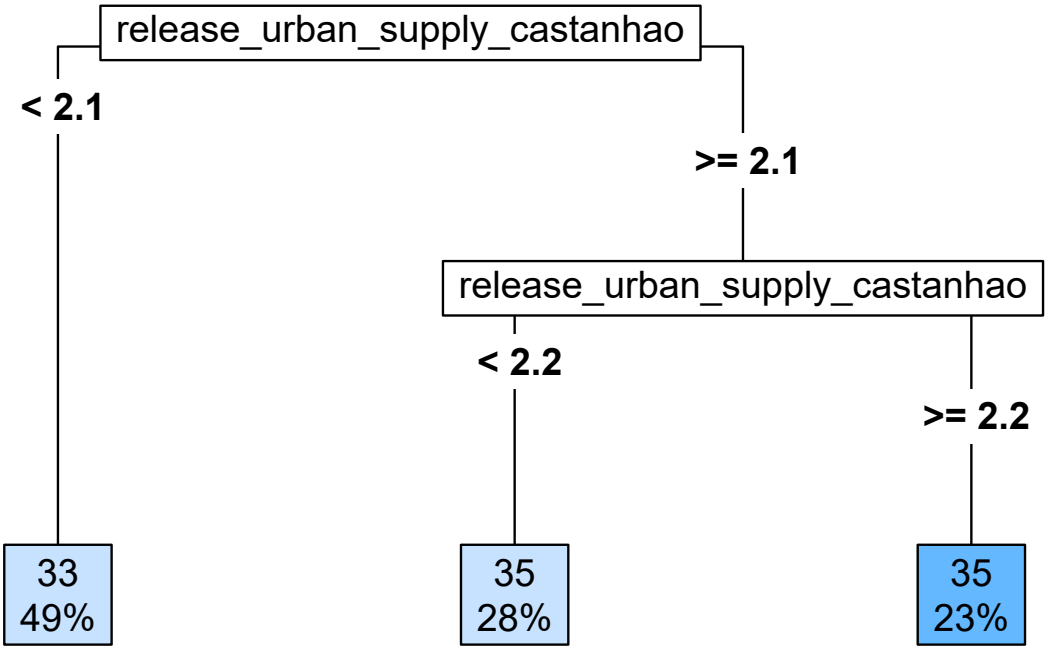
Response: Release − Irrigation Castanhão
Month: January



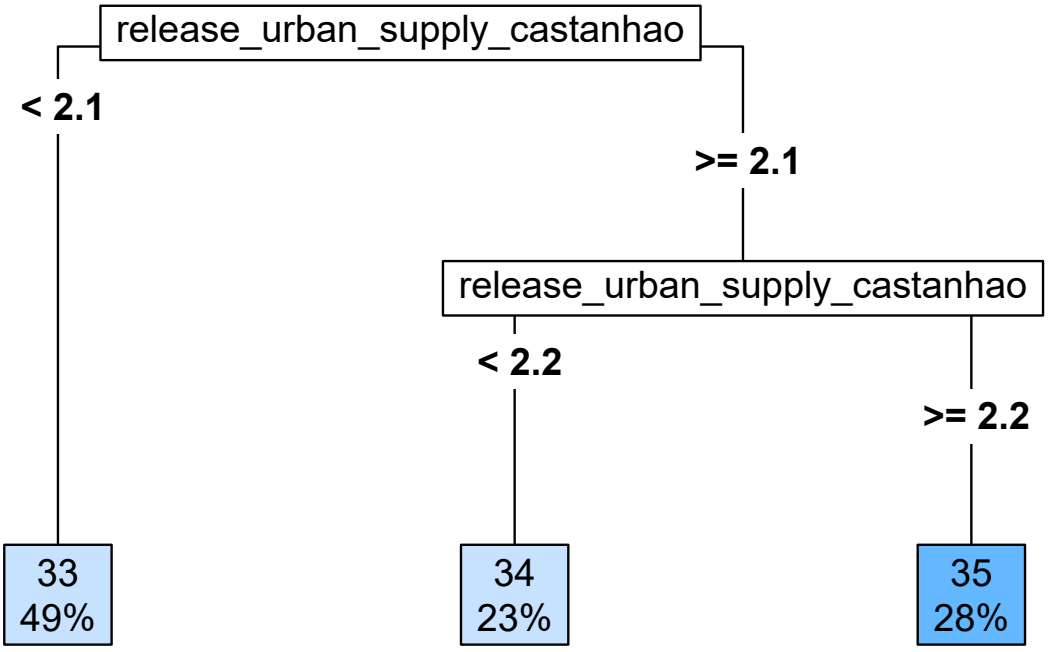
Response: Release − Irrigation Castanhão
Month: February



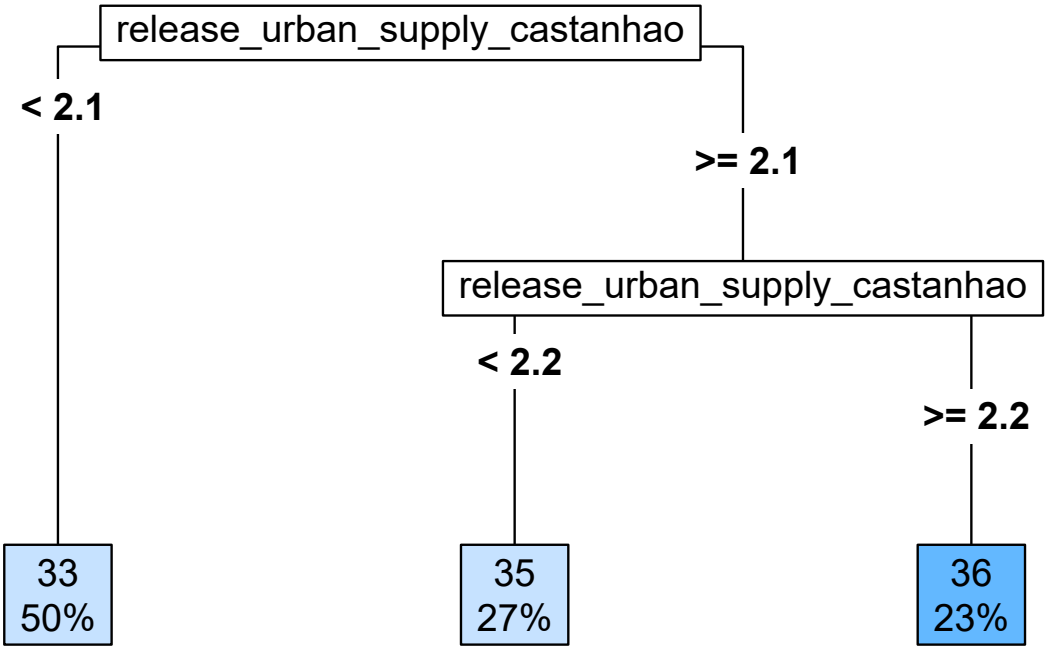
Response: Release − Irrigation Castanhão
Month: March



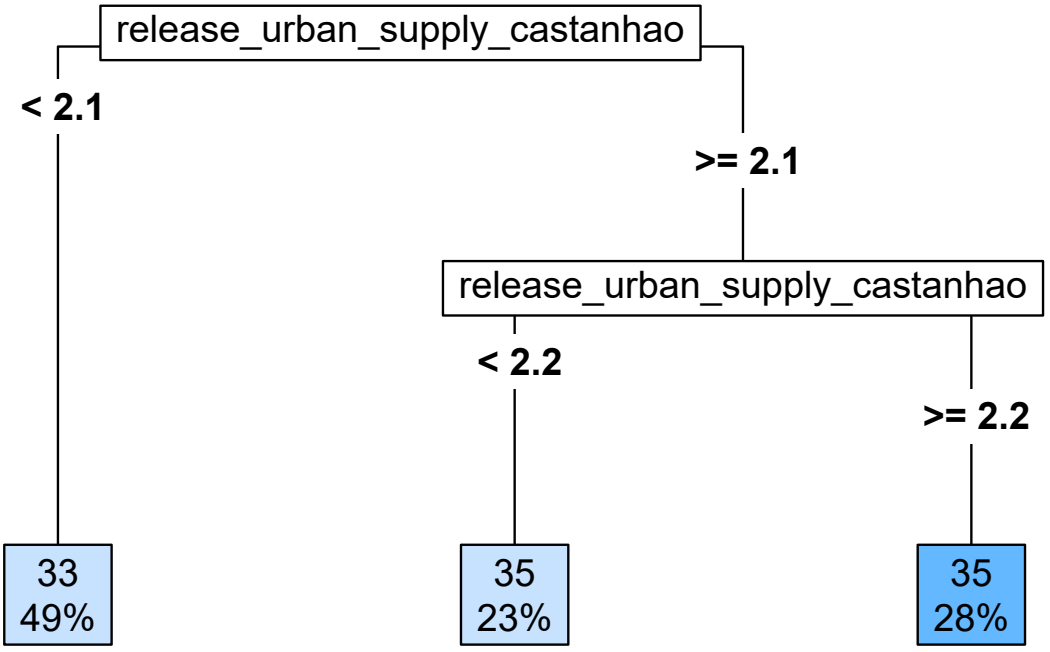
Response: Release − Irrigation Castanhão
Month: April



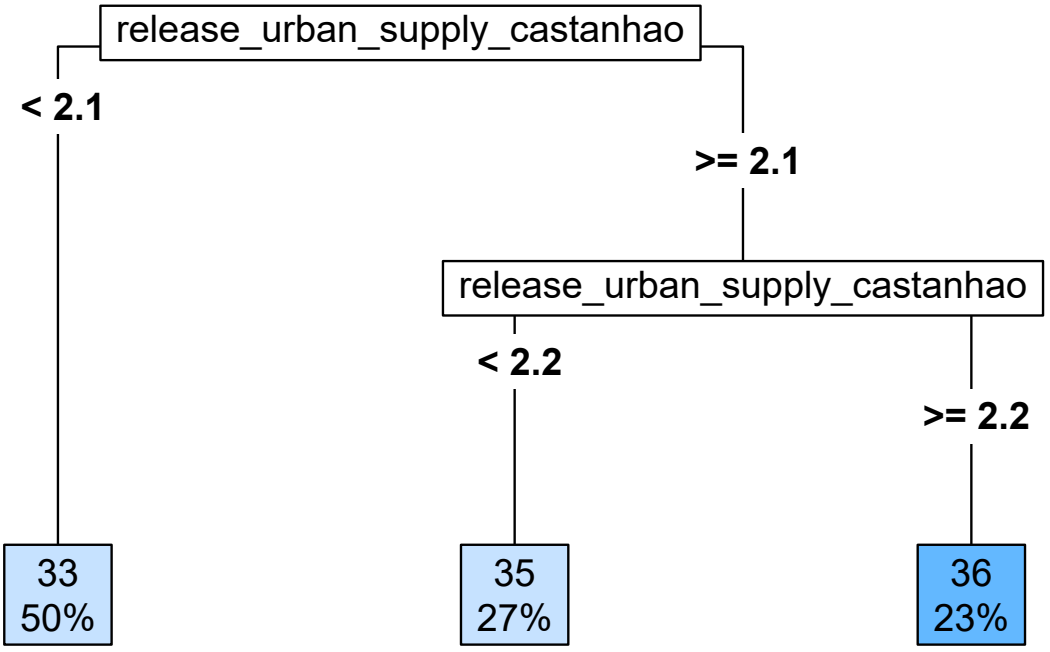
Response: Release − Irrigation Castanhão
Month: May



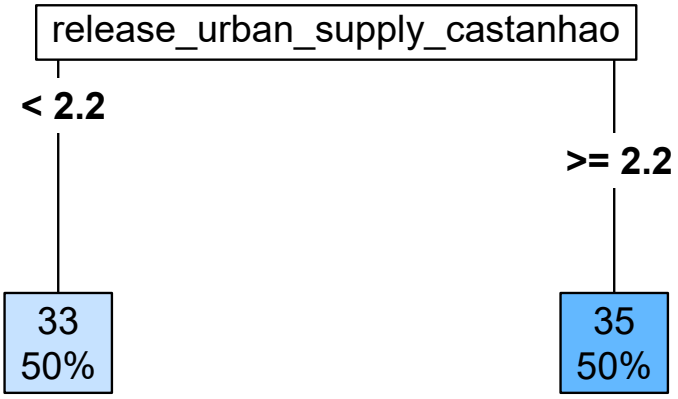
Response: Release − Irrigation Castanhão
Month: June



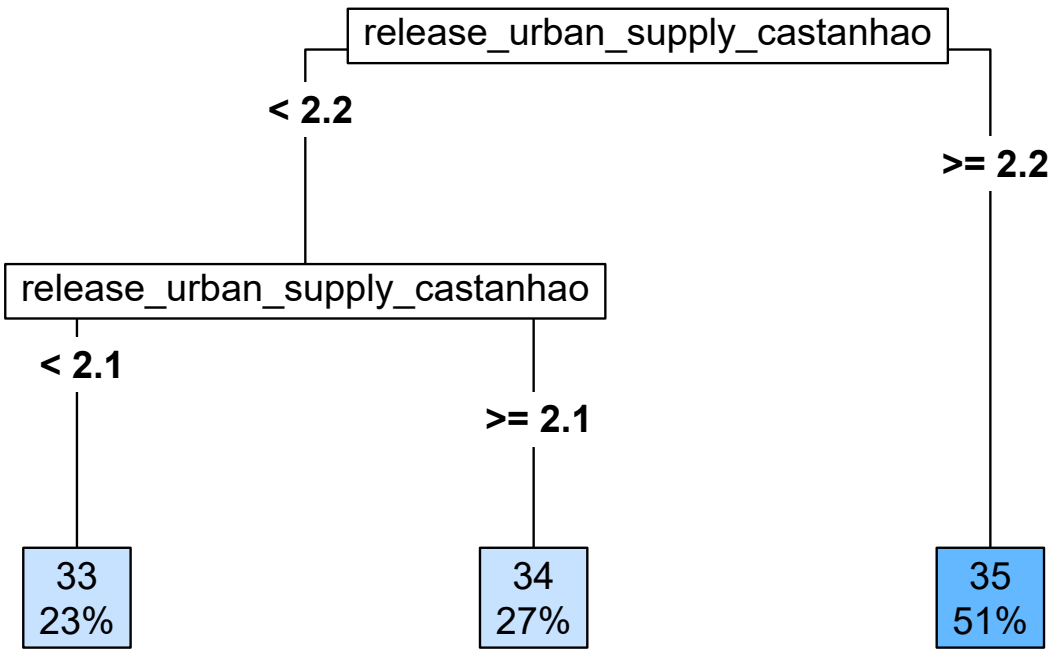
Response: Release − Irrigation Castanhão
Month: July



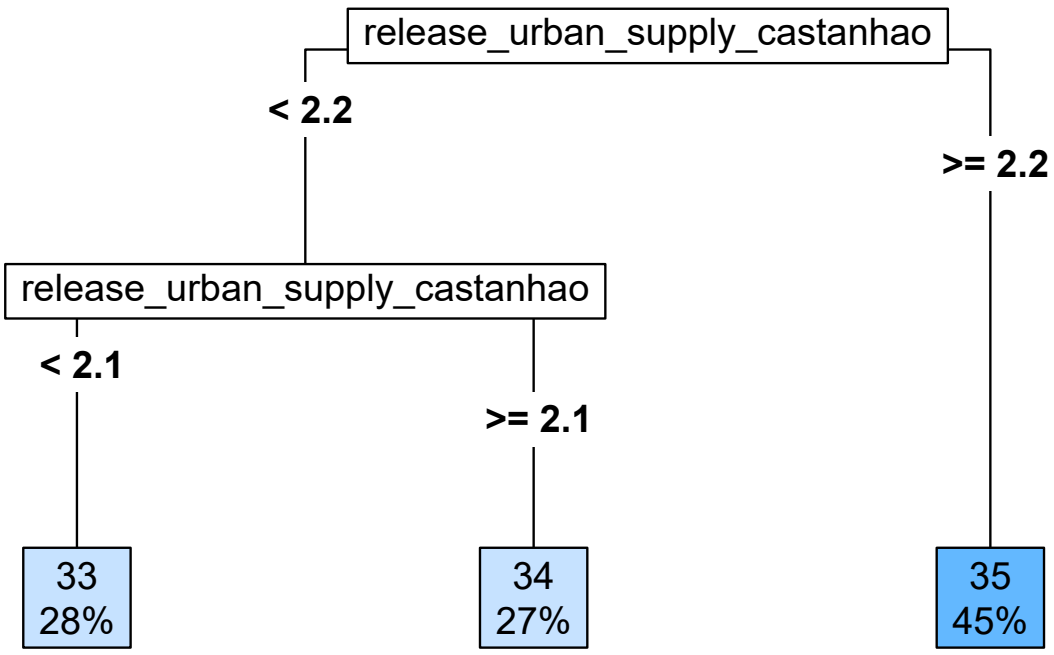
Response: Release − Irrigation Castanhão
Month: August



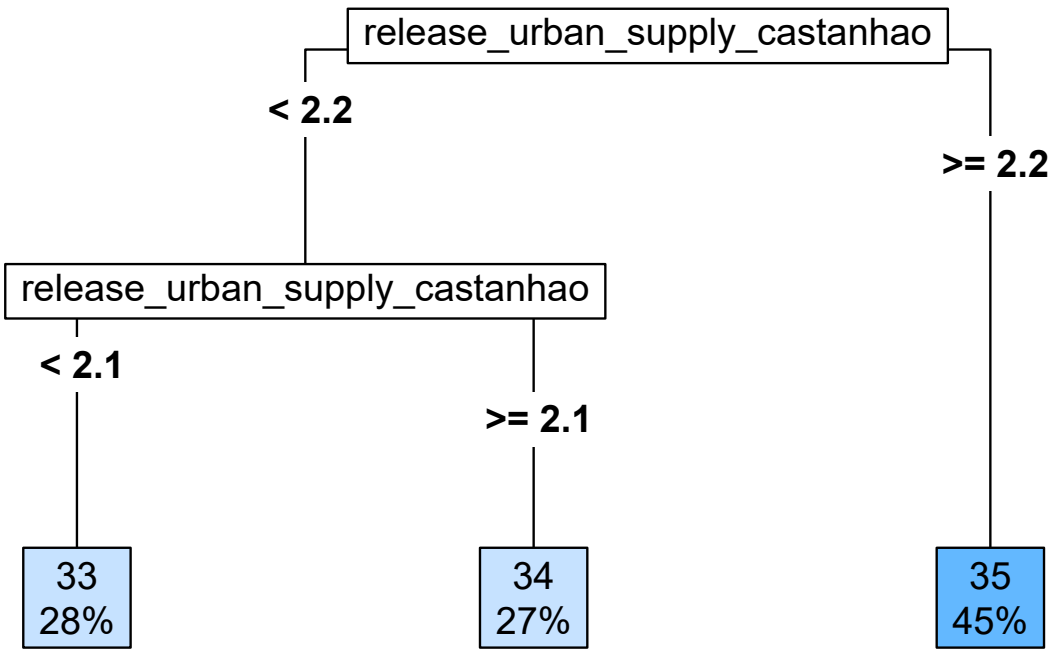
Response: Release − Irrigation Castanhão
Month: September



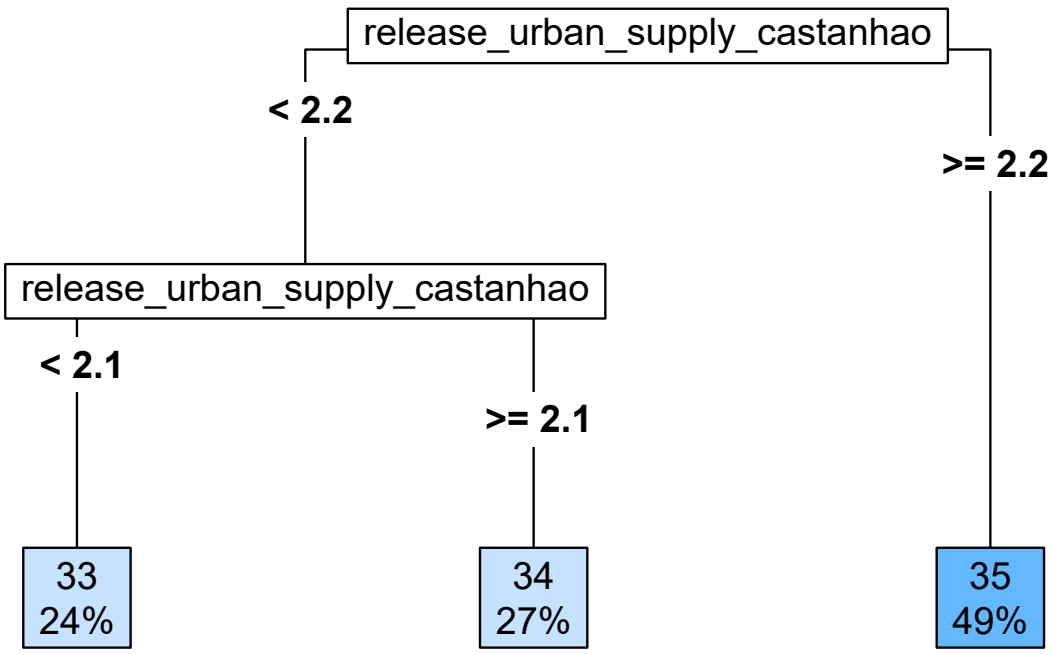
Response: Release − Irrigation Castanhão
Month: October

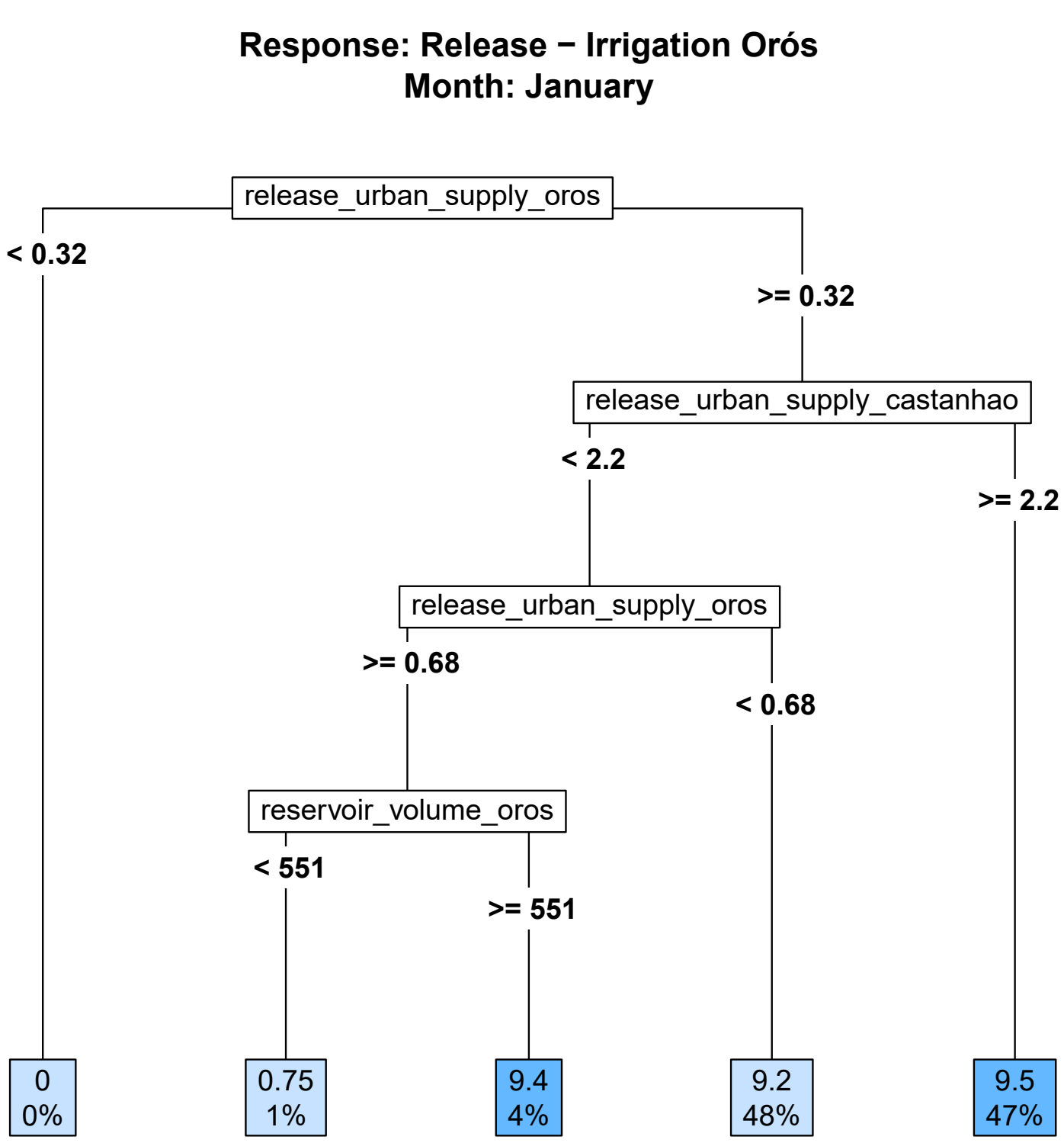


Response: Release − Irrigation Castanhão
Month: November

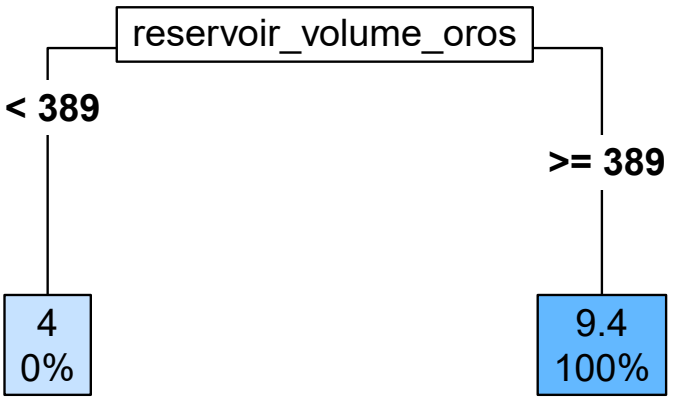


Response: Release − Irrigation Castanhão
Month: December

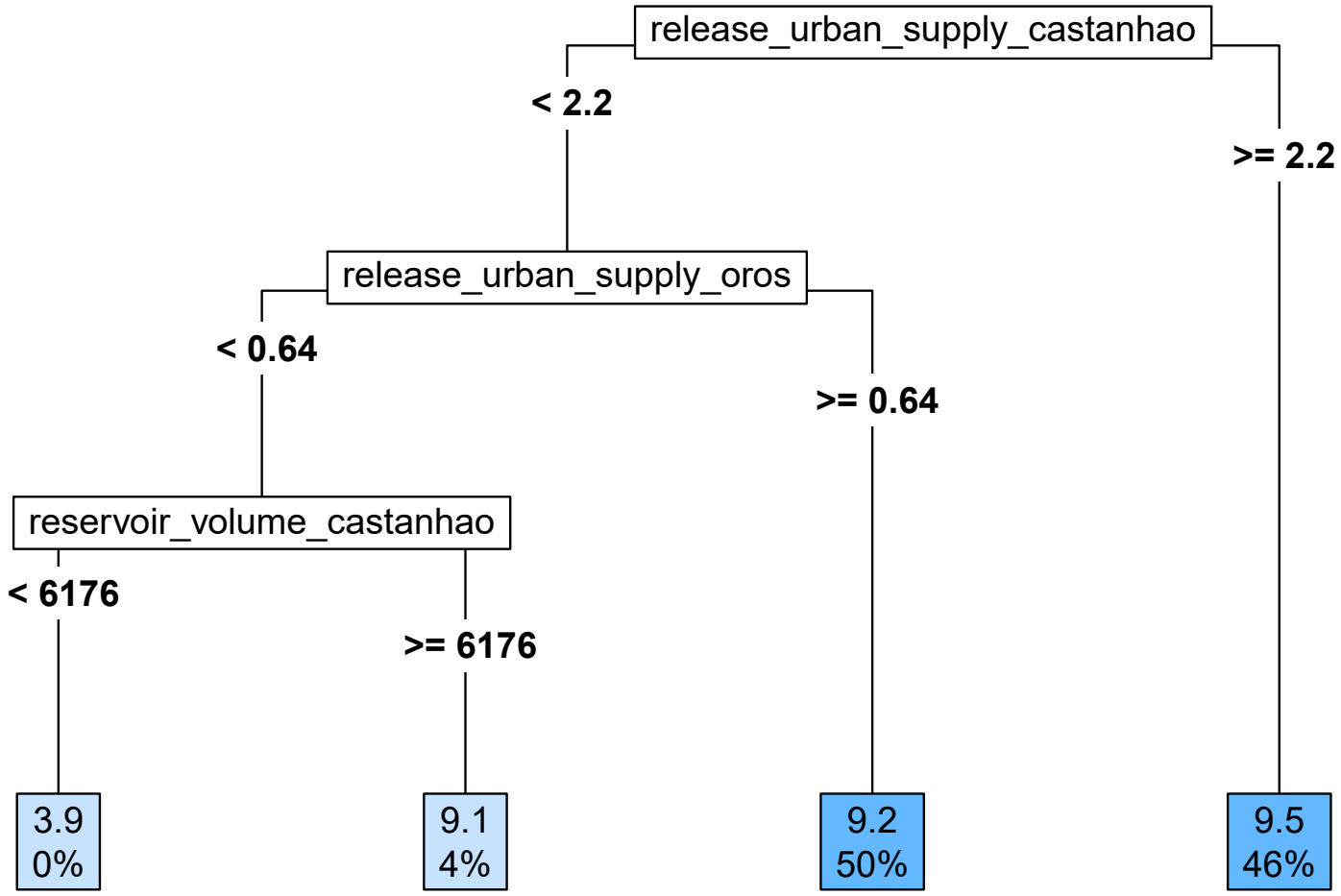



Response: Release − Irrigation Orós
Month: January
release_urban_supply_oros
< 0.32
>= 0.32
release_urban_supply_castanhao
< 2.2
>= 2.2
release_urban_supply_oros
>= 0.68
< 0.68
reservoir_volume_oros
< 551
>= 551
0
0%
0.75
1%
9.4
4%
9.2
48%
9.5
47%

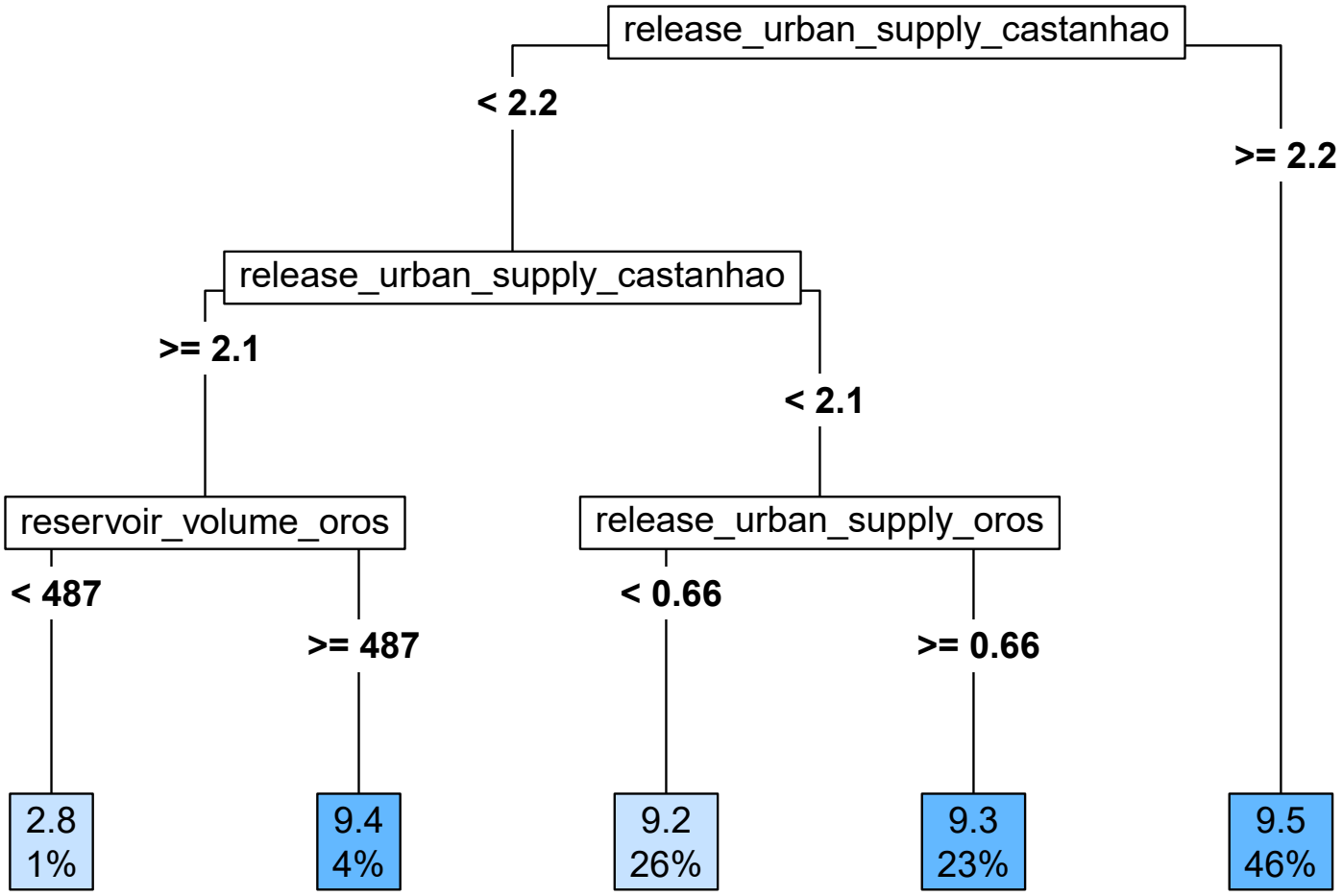
Response: Release − Irrigation Orós
Month: February



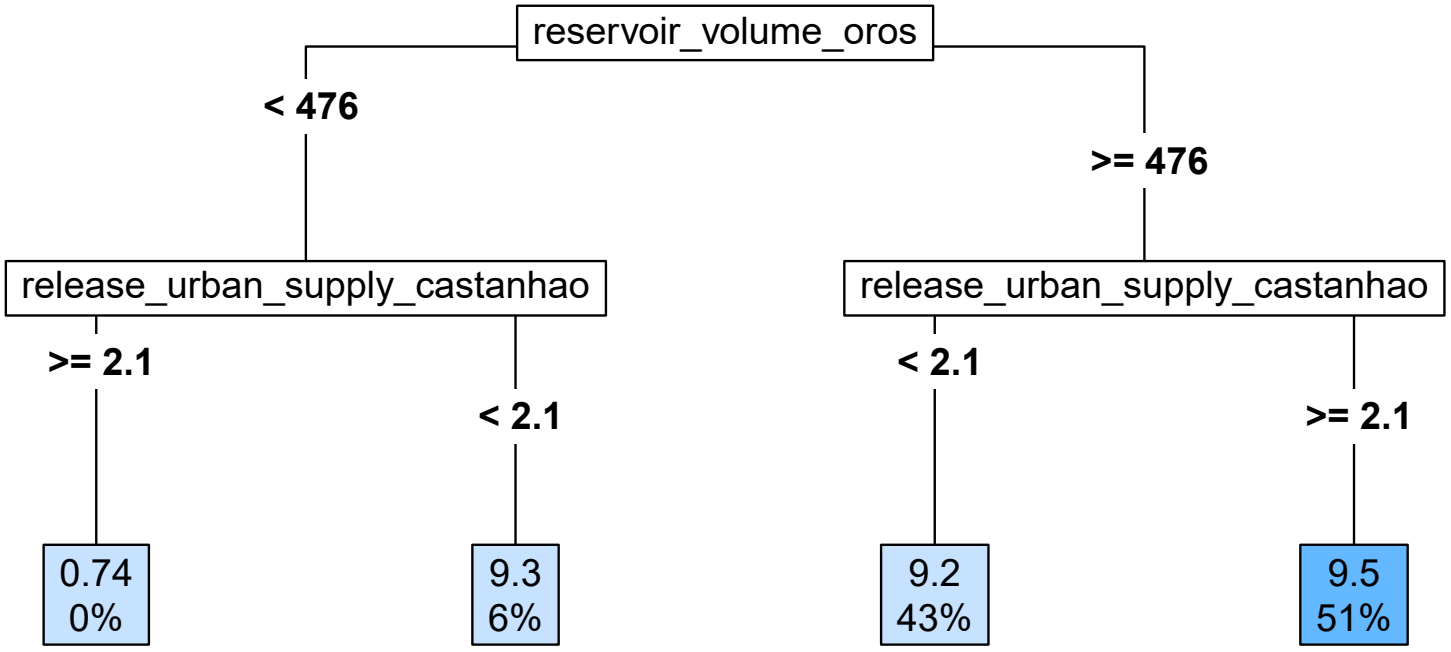
Response: Release − Irrigation Orós
Month: March



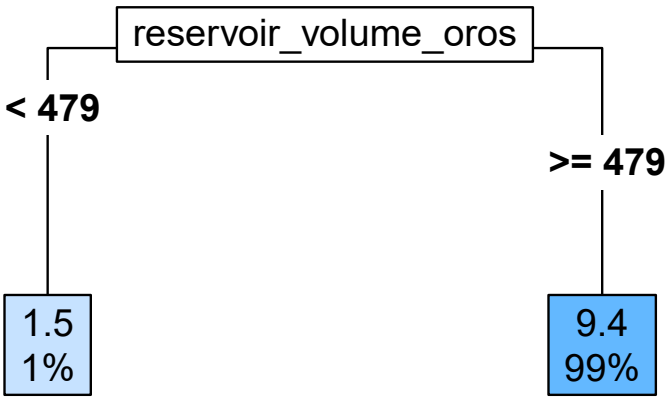
Response: Release − Irrigation Orós
Month: April



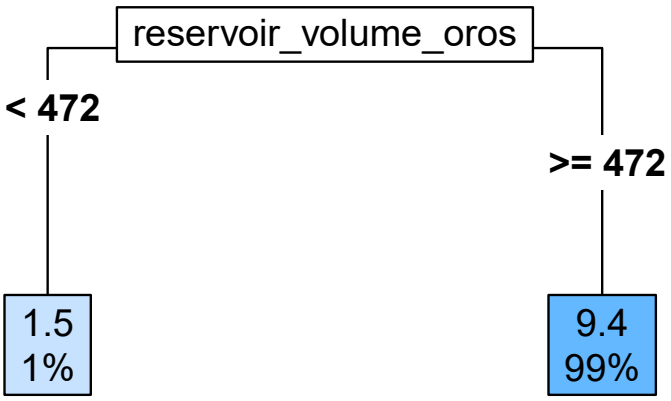
Response: Release − Irrigation Orós
Month: May



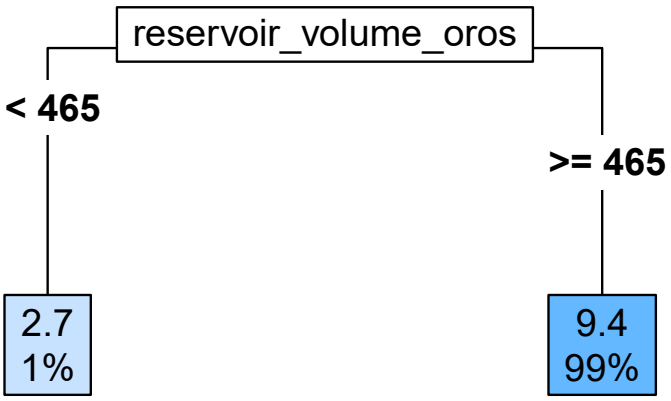
Response: Release − Irrigation Orós
Month: June



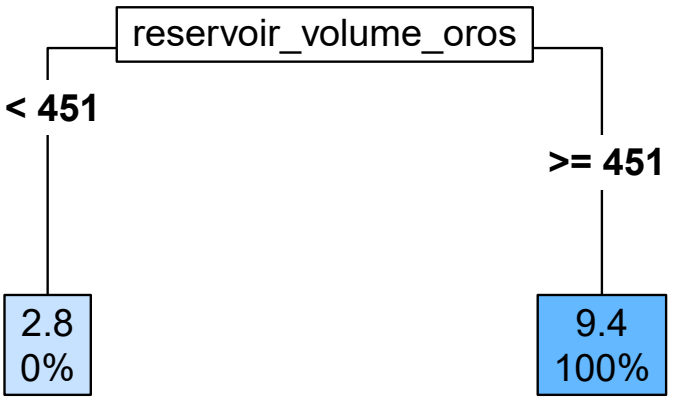
Response: Release − Irrigation Orós
Month: July



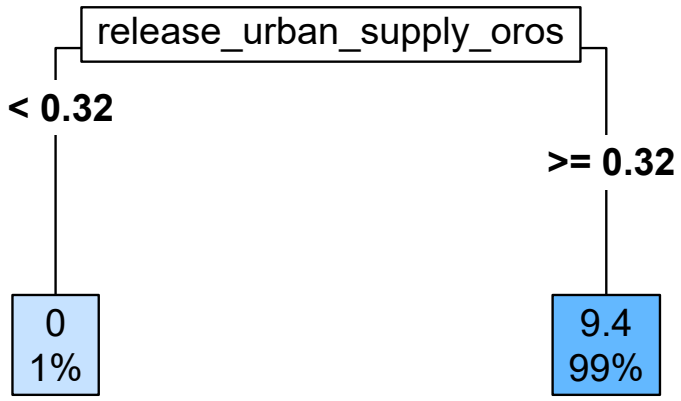
Response: Release − Irrigation Orós
Month: August



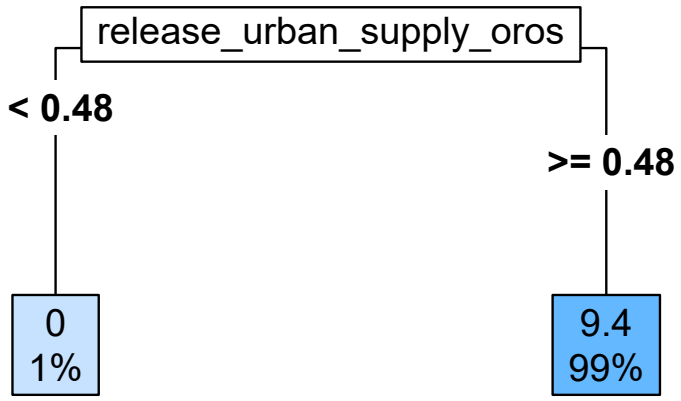
Response: Release − Irrigation Orós
Month: September



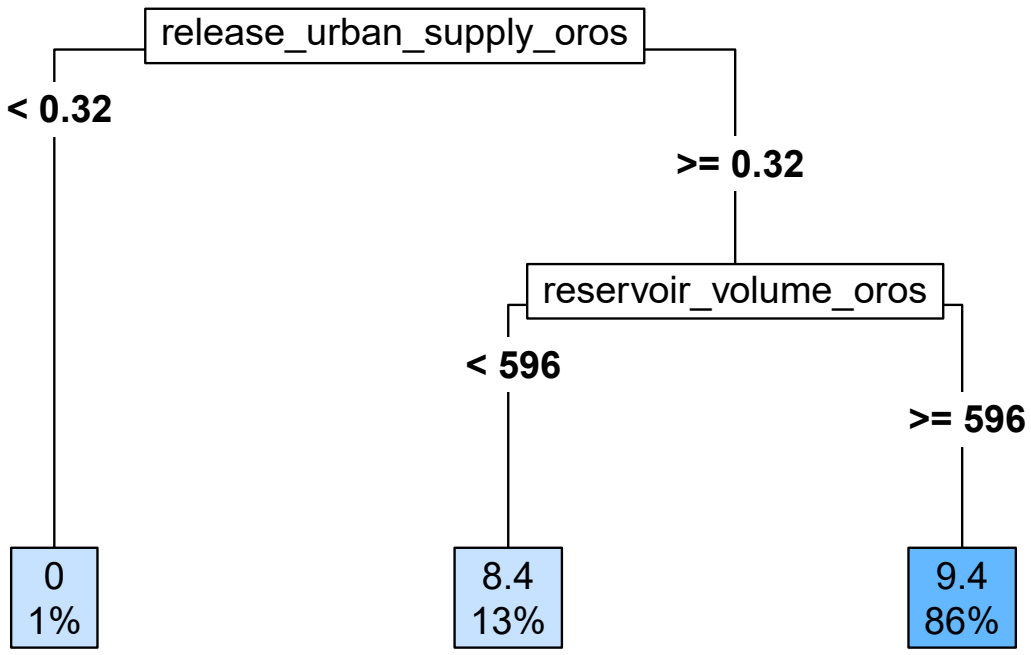
Response: Release − Irrigation Orós
Month: October



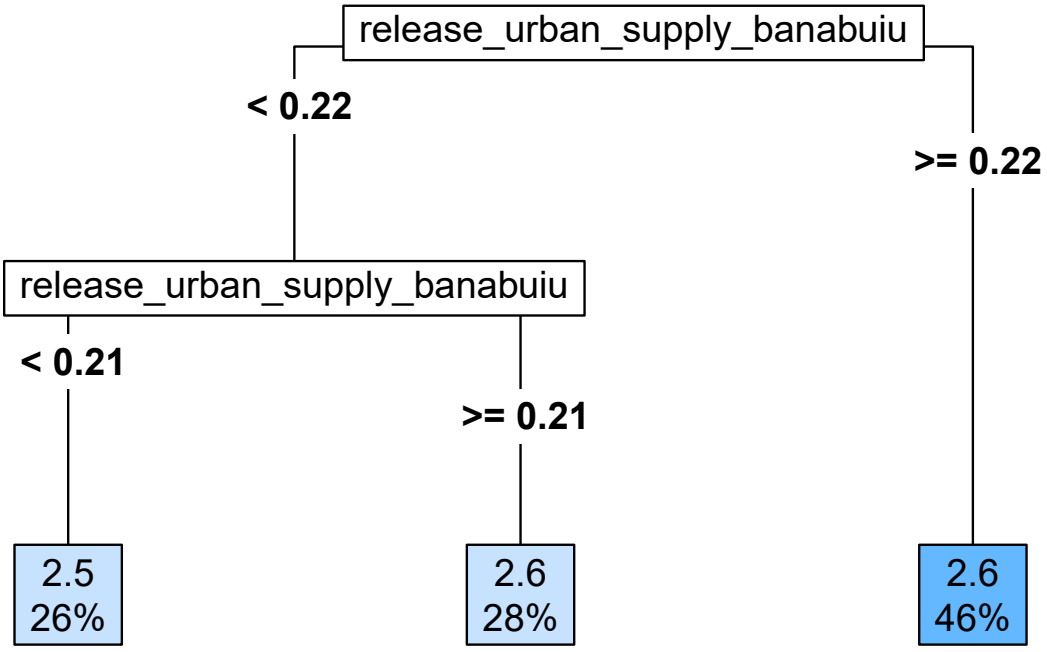
Response: Release − Irrigation Orós
Month: November



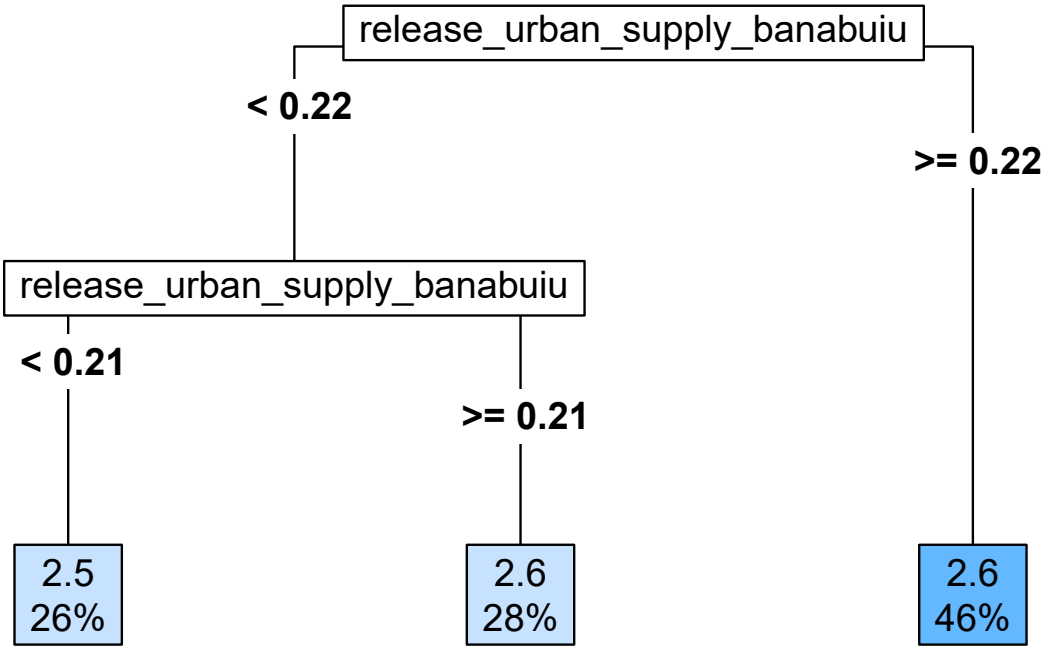
Response: Release − Irrigation Orós
Month: December



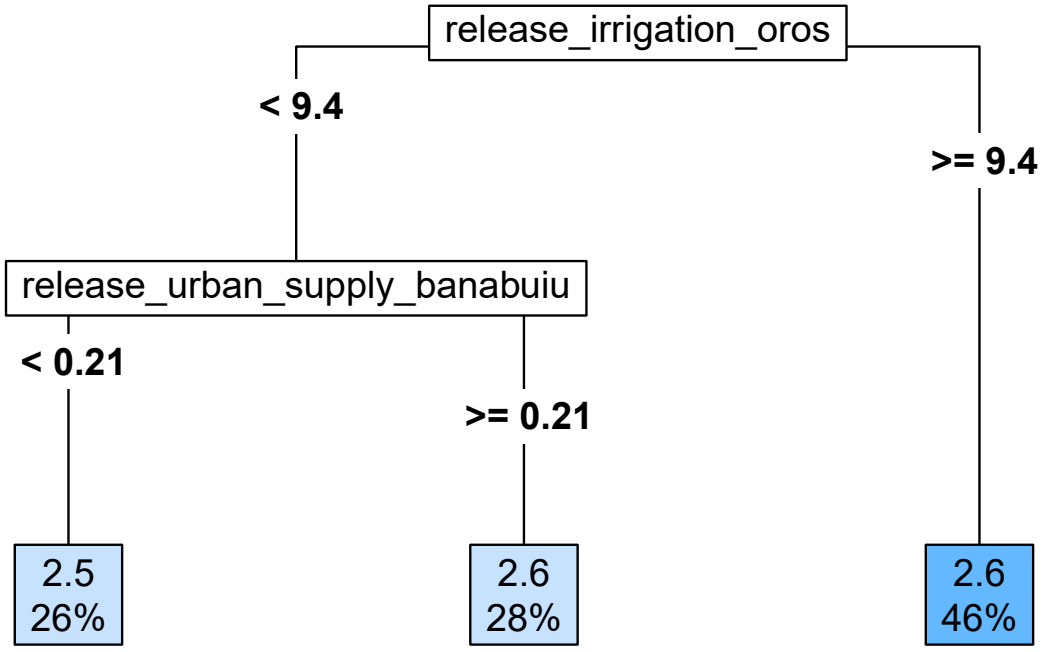
Response: Release − Irrigation Banabuiú
Month: January



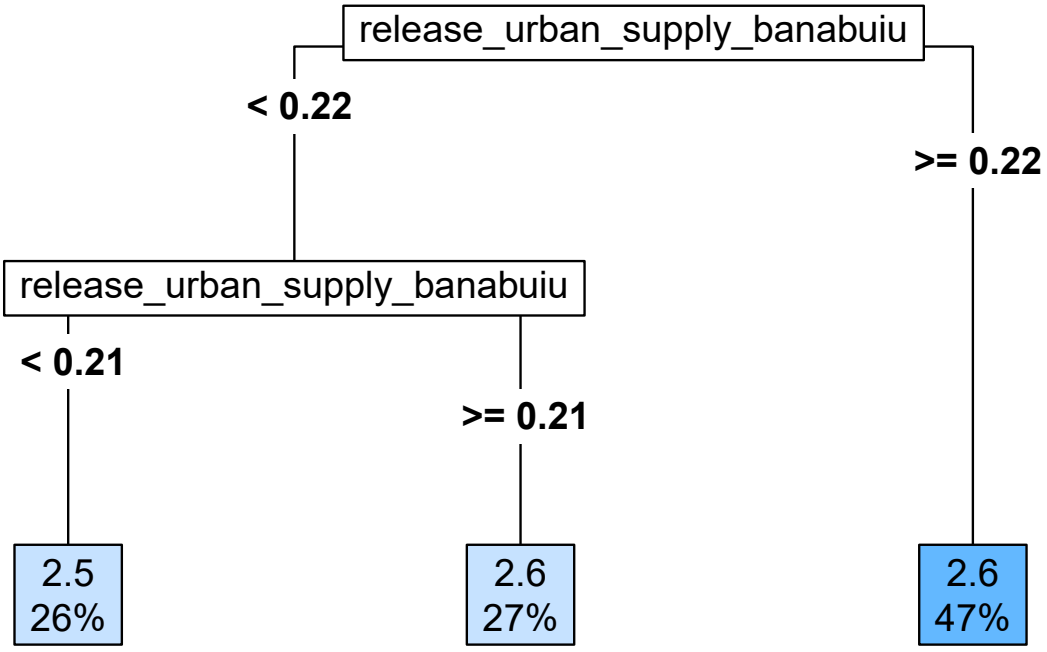
Response: Release − Irrigation Banabuiú
Month: February



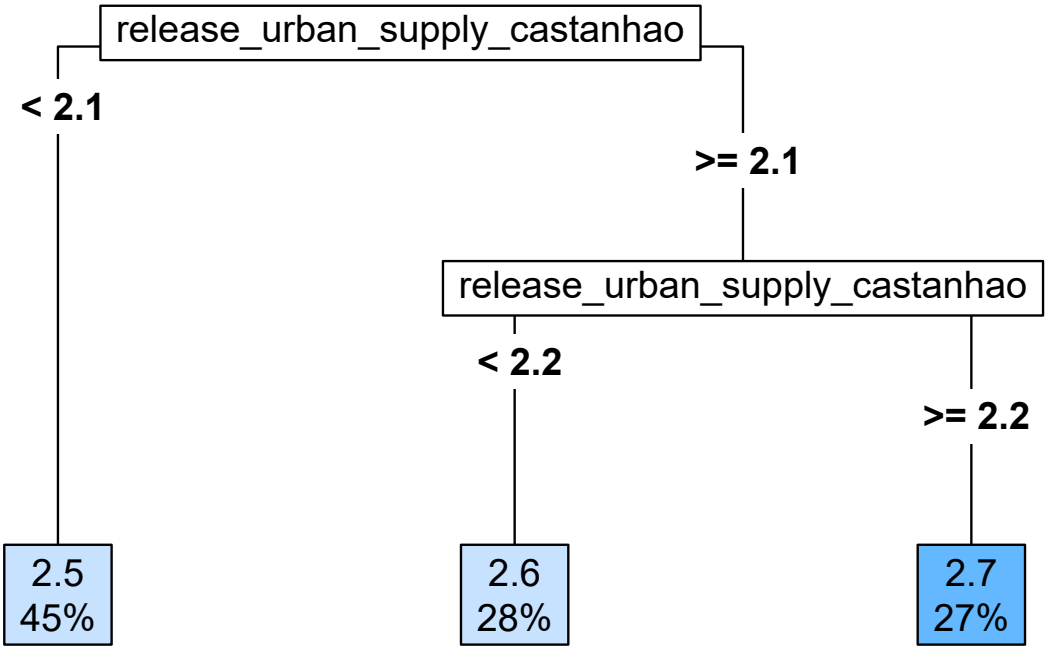
Response: Release − Irrigation Banabuiú
Month: March



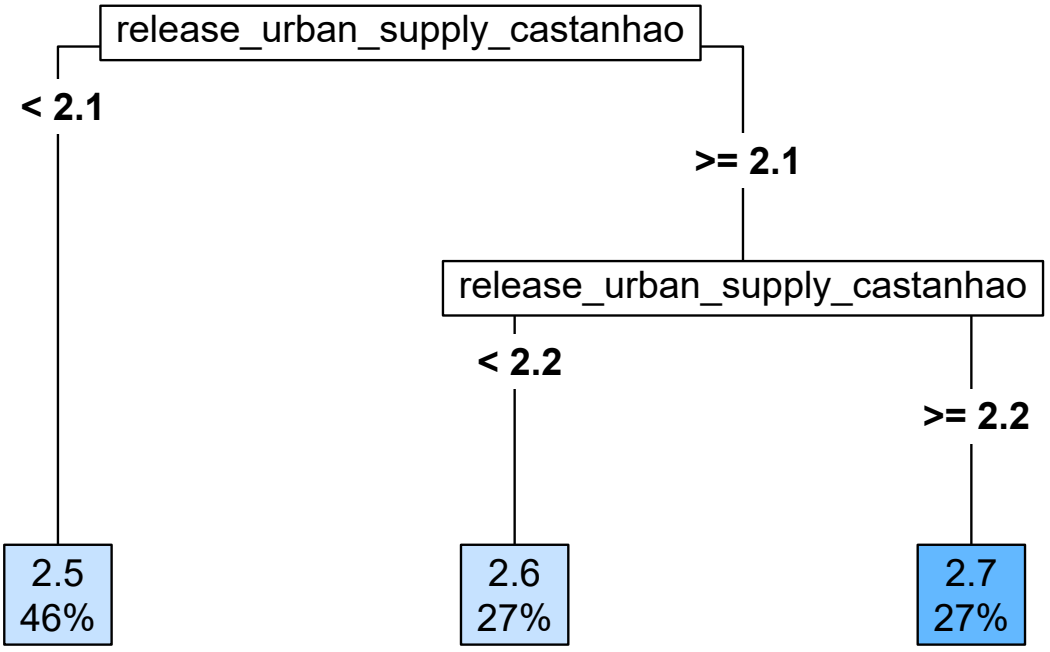
Response: Release − Irrigation Banabuiú
Month: April



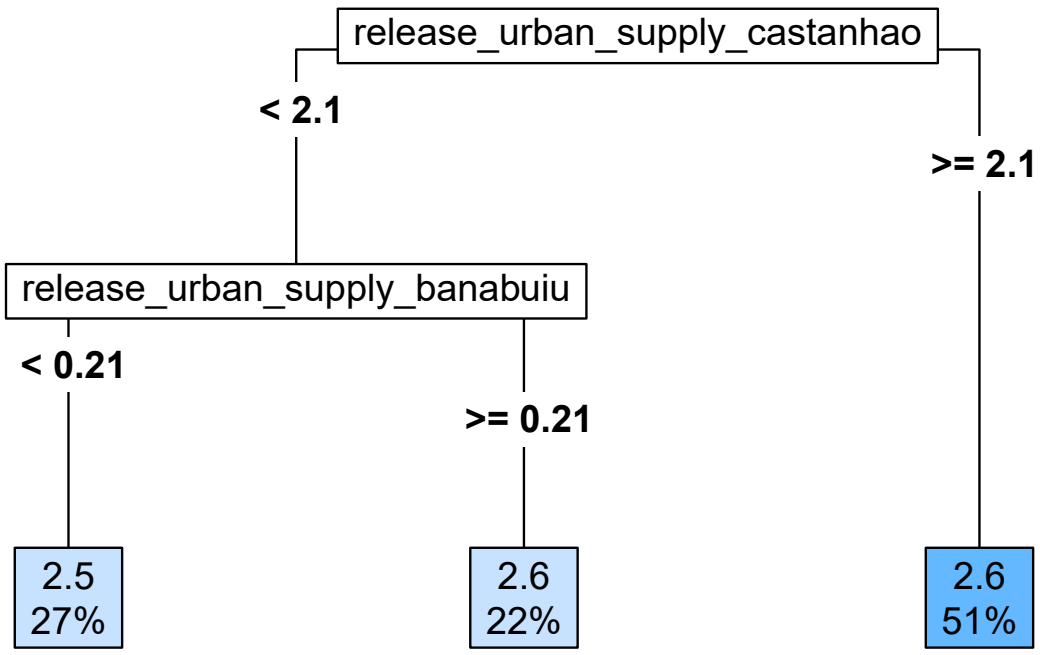
Response: Release − Irrigation Banabuiú
Month: May



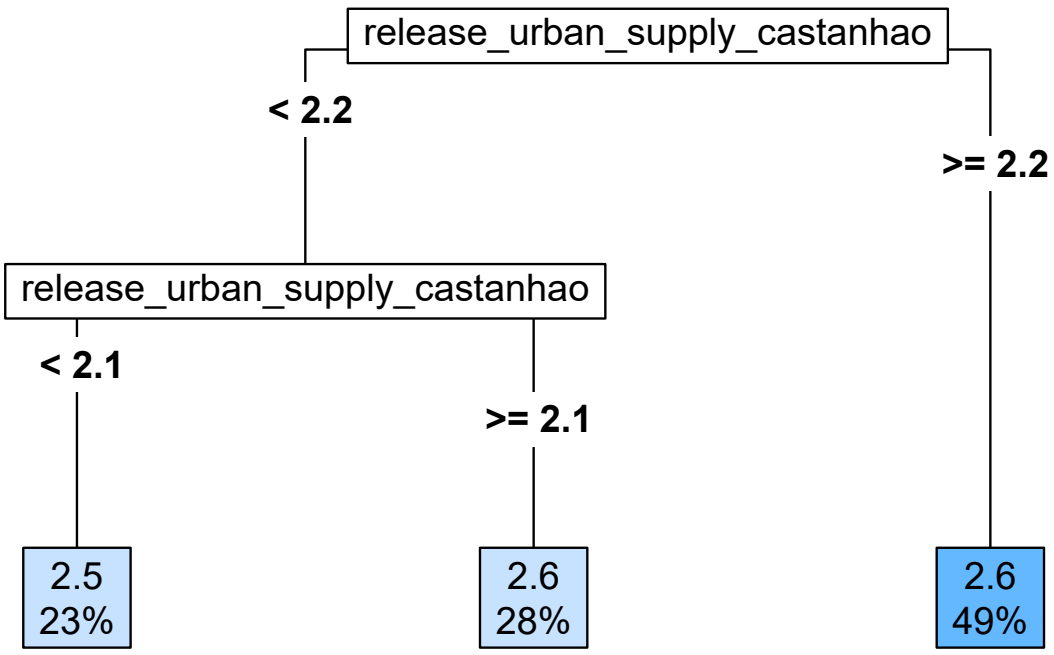
Response: Release − Irrigation Banabuiú
Month: June



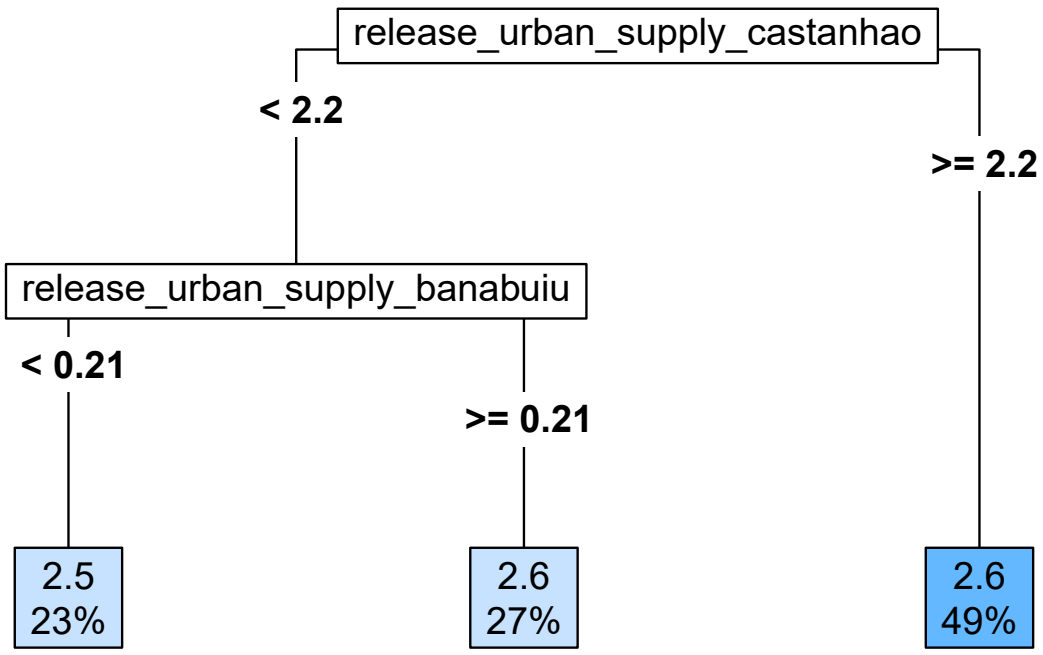
Response: Release − Irrigation Banabuiú
Month: July



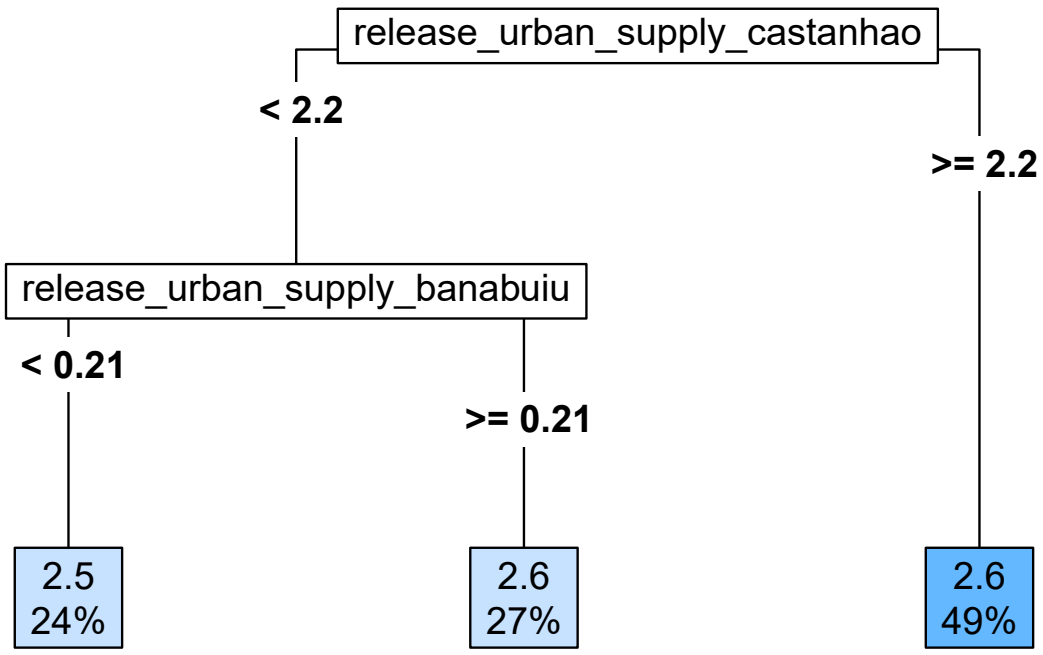
Response: Release − Irrigation Banabuiú
Month: August



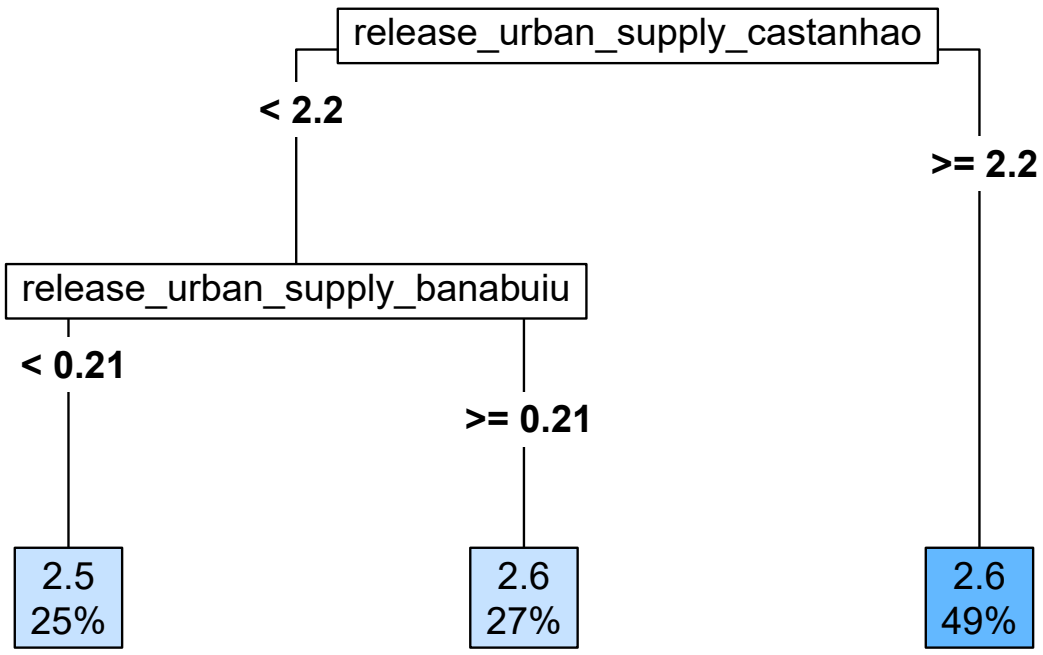
Response: Release − Irrigation Banabuiú
Month: September



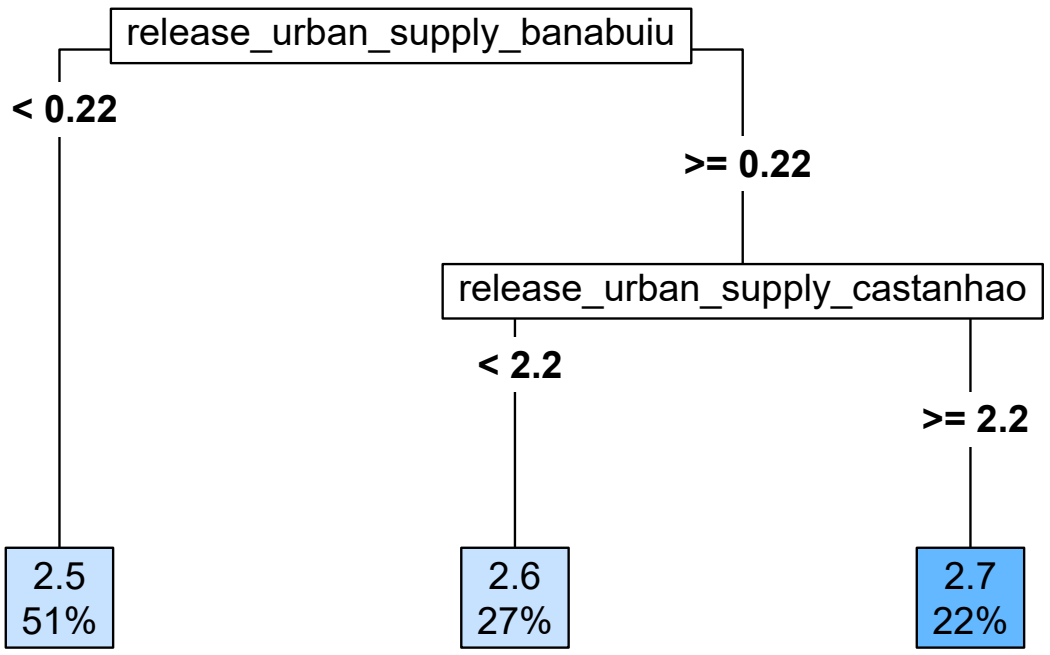
Response: Release − Irrigation Banabuiú
Month: October



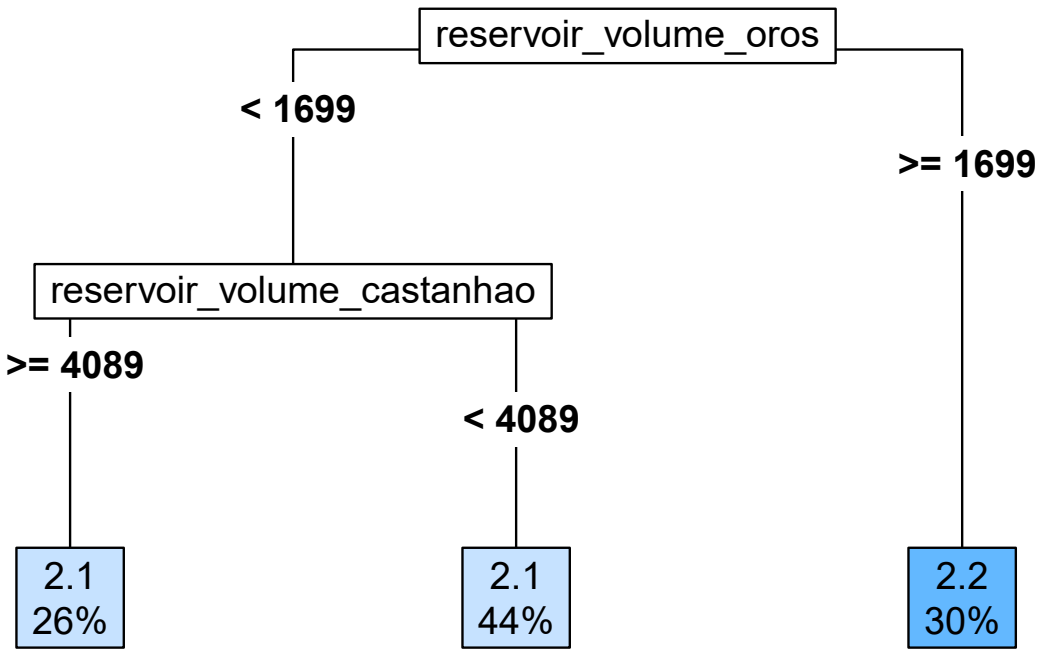
Response: Release − Irrigation Banabuiú
Month: November



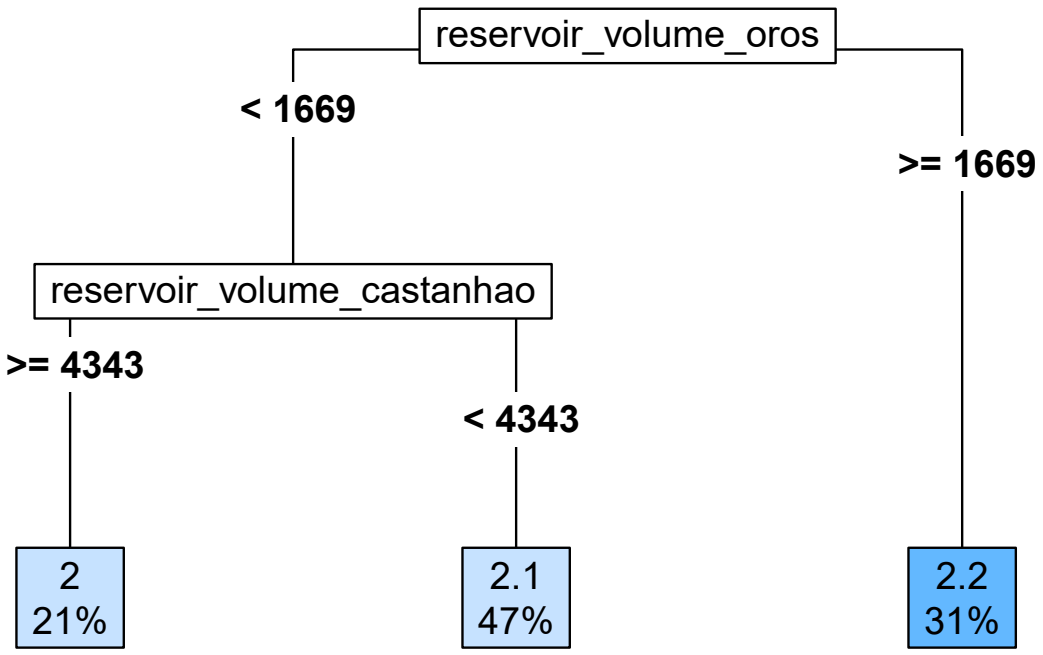
Response: Release − Irrigation Banabuiú
Month: December



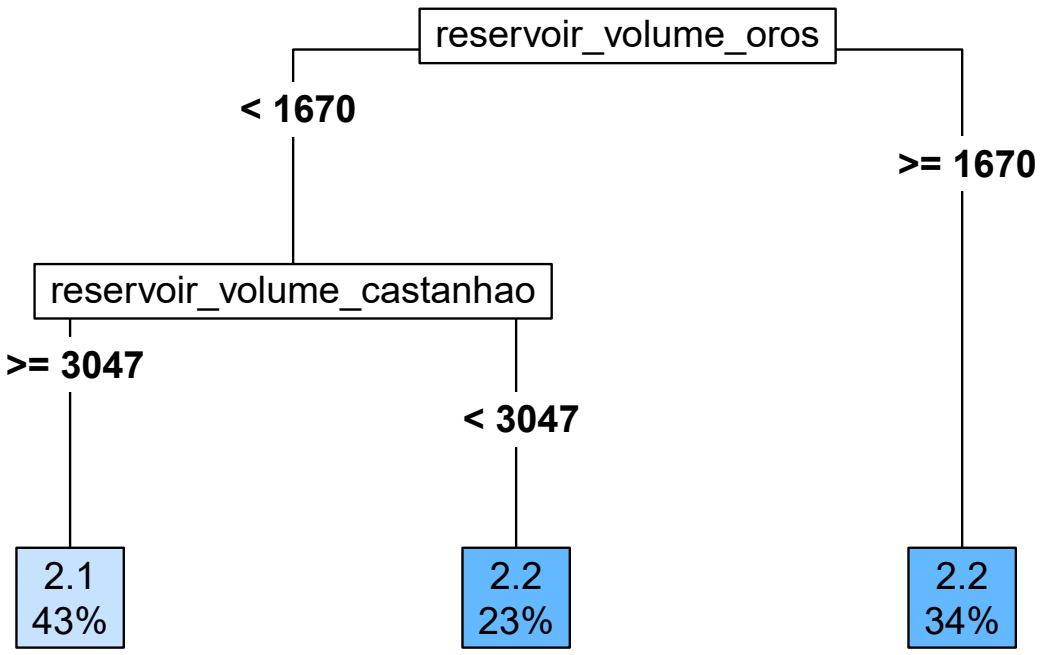
Response: Release for Urban supply − Castanhão
Month: January



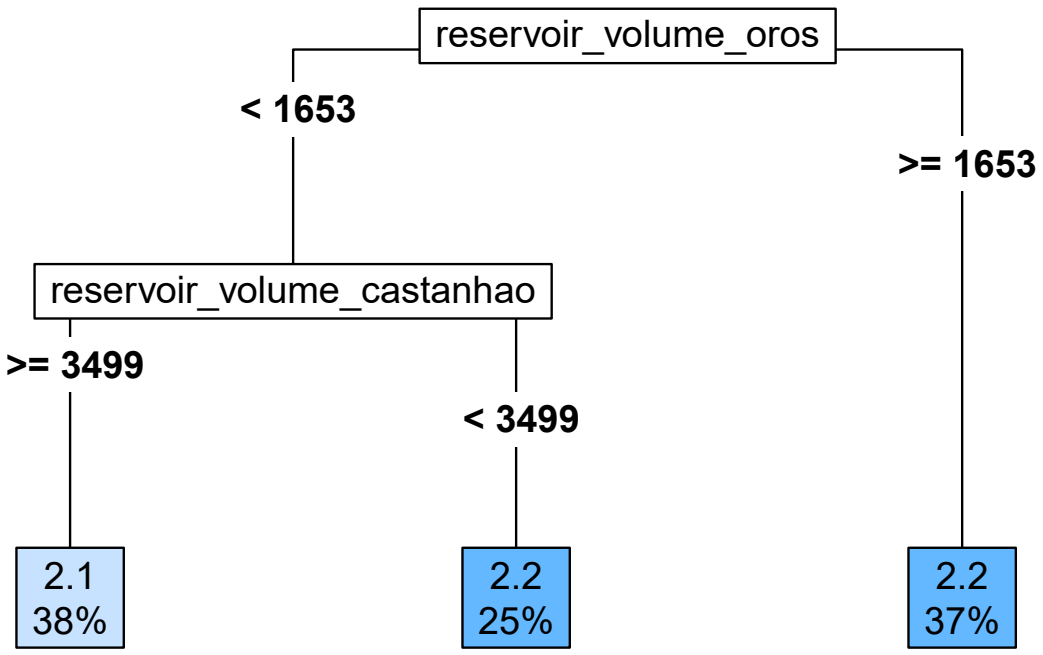
Response: Release for Urban supply − Castanhão
Month: February



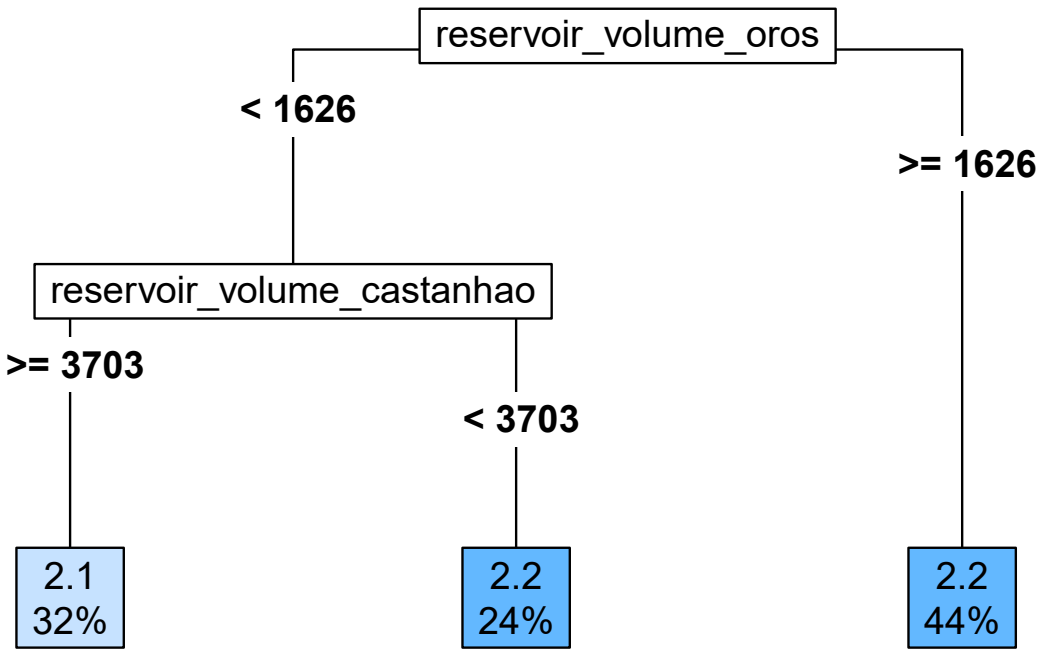
Response: Release for Urban supply − Castanhão
Month: March



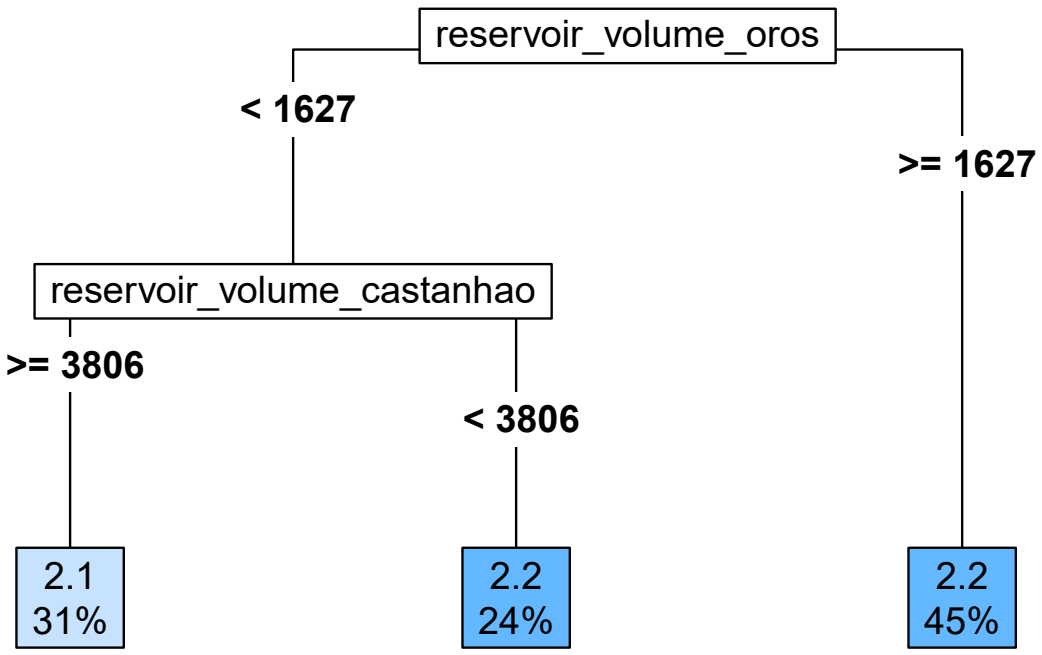
Response: Release for Urban supply − Castanhão
Month: April



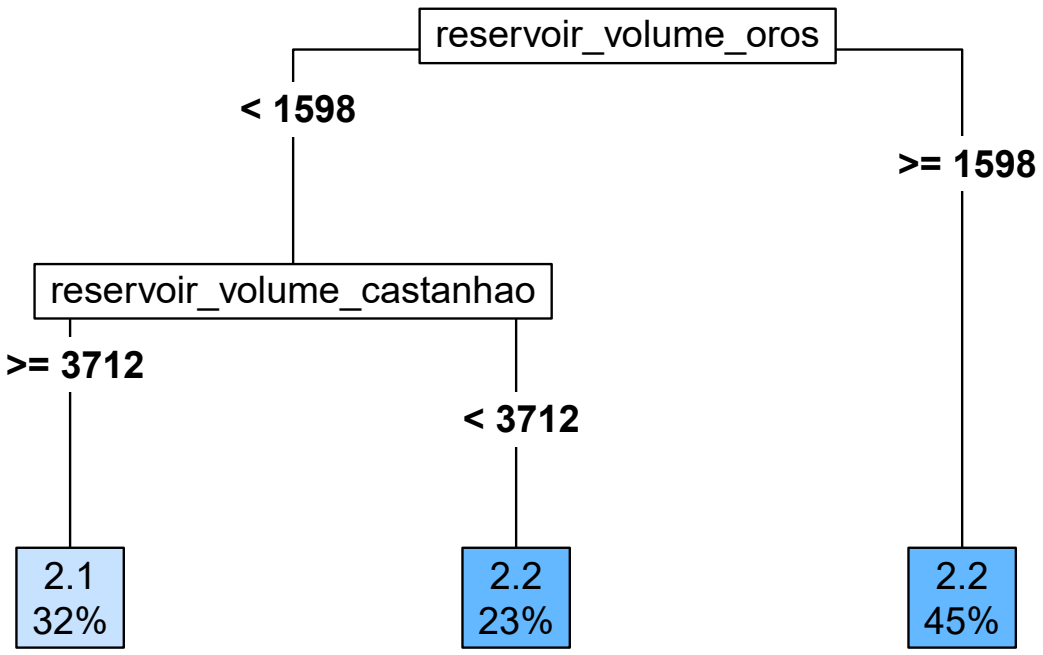
Response: Release for Urban supply − Castanhão
Month: May



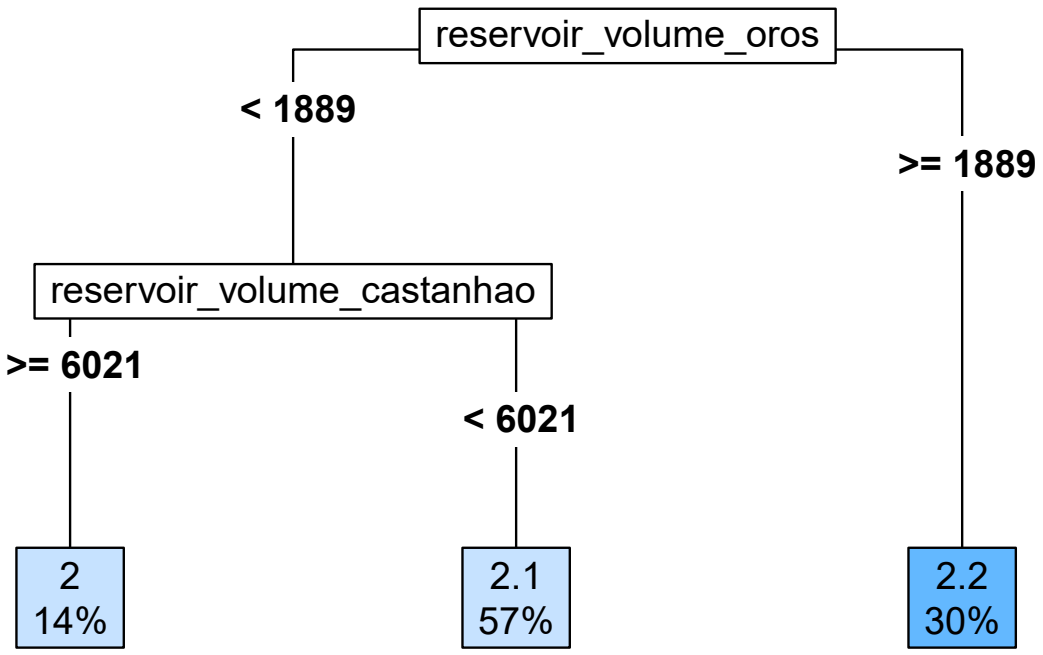
Response: Release for Urban supply − Castanhão
Month: June



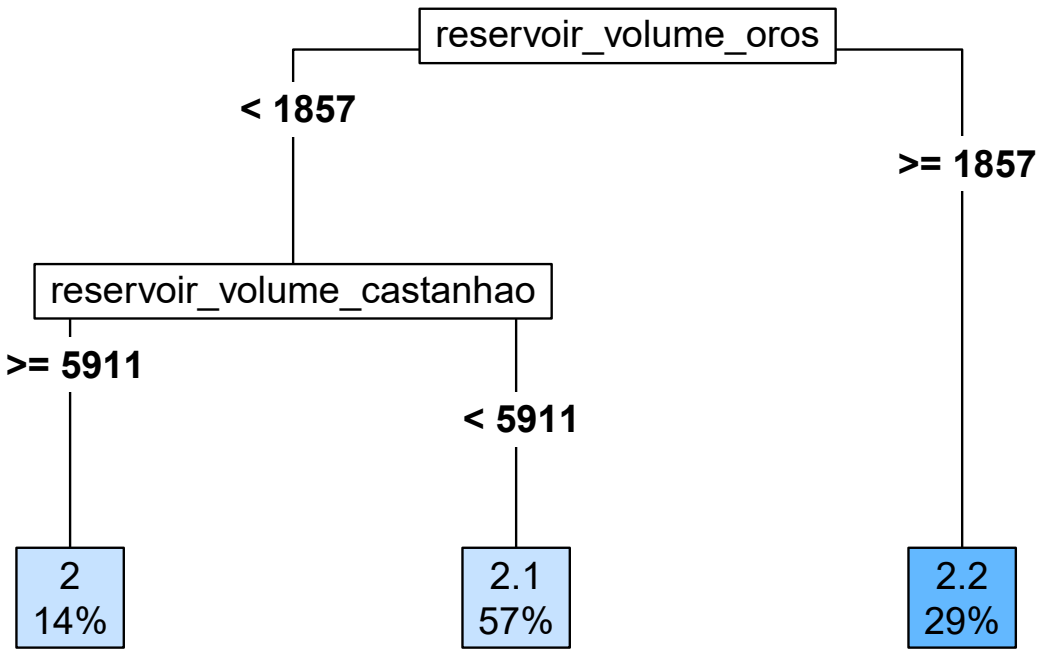
Response: Release for Urban supply − Castanhão
Month: July



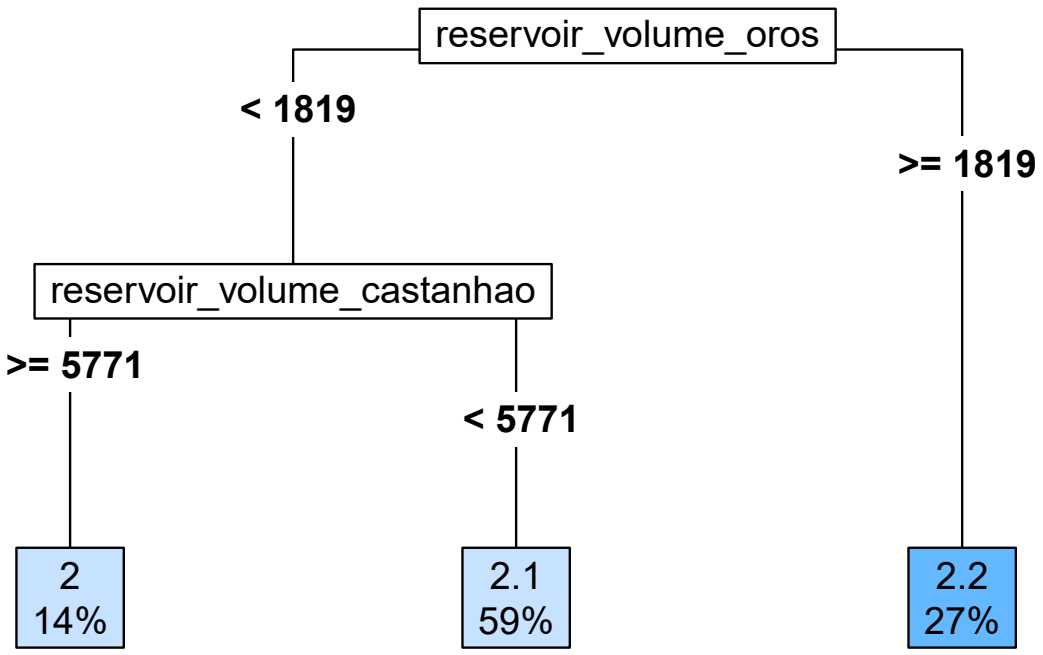
Response: Release for Urban supply − Castanhão
Month: August



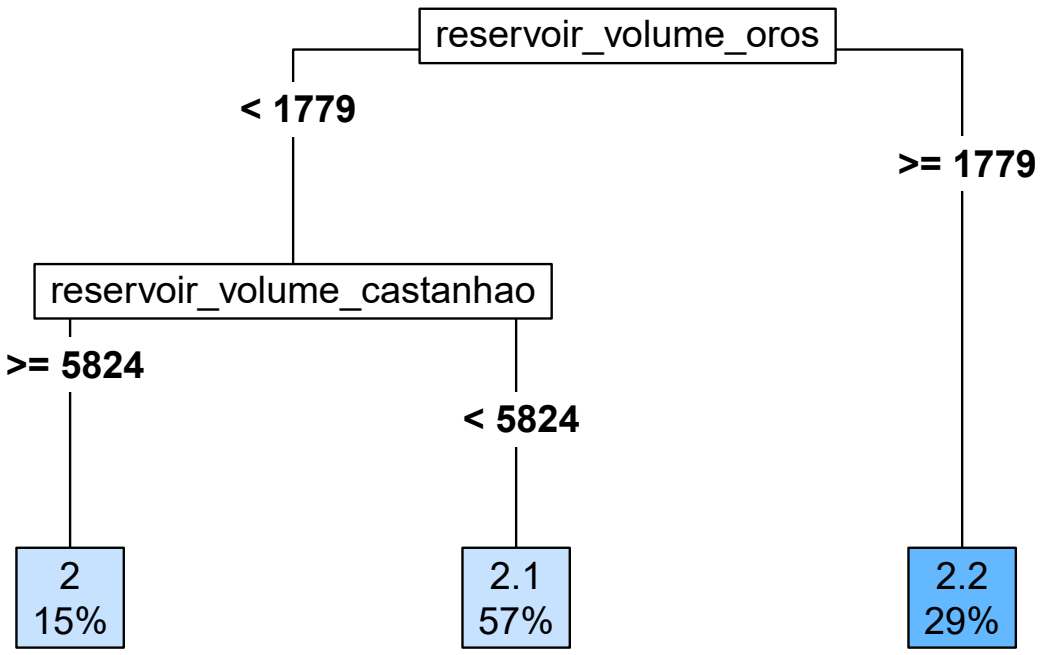
Response: Release for Urban supply − Castanhão
Month: September



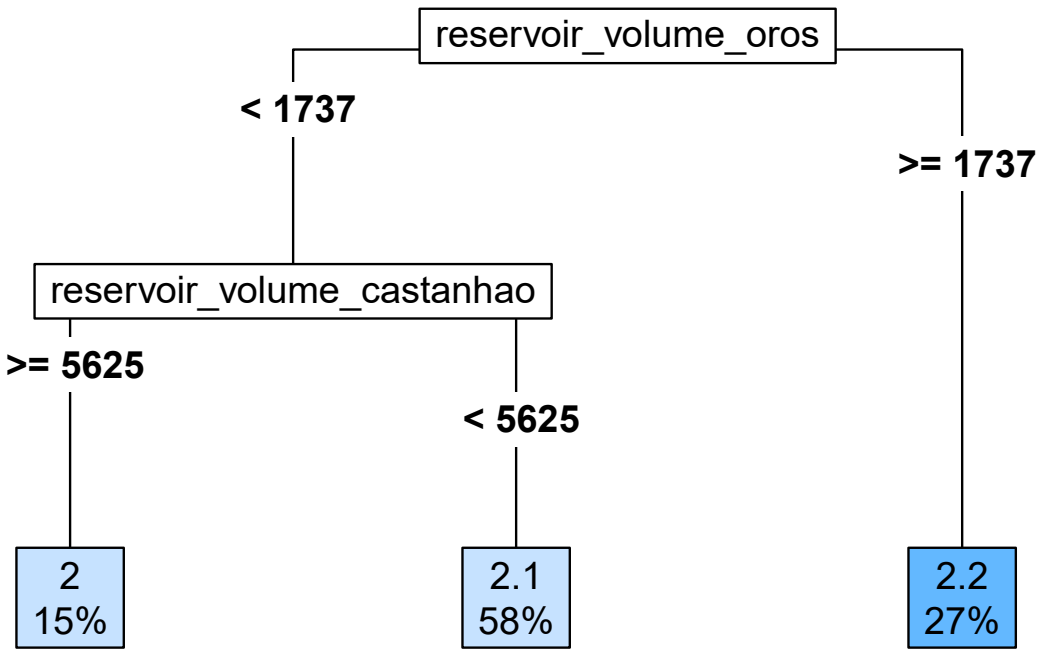
Response: Release for Urban supply − Castanhão
Month: October



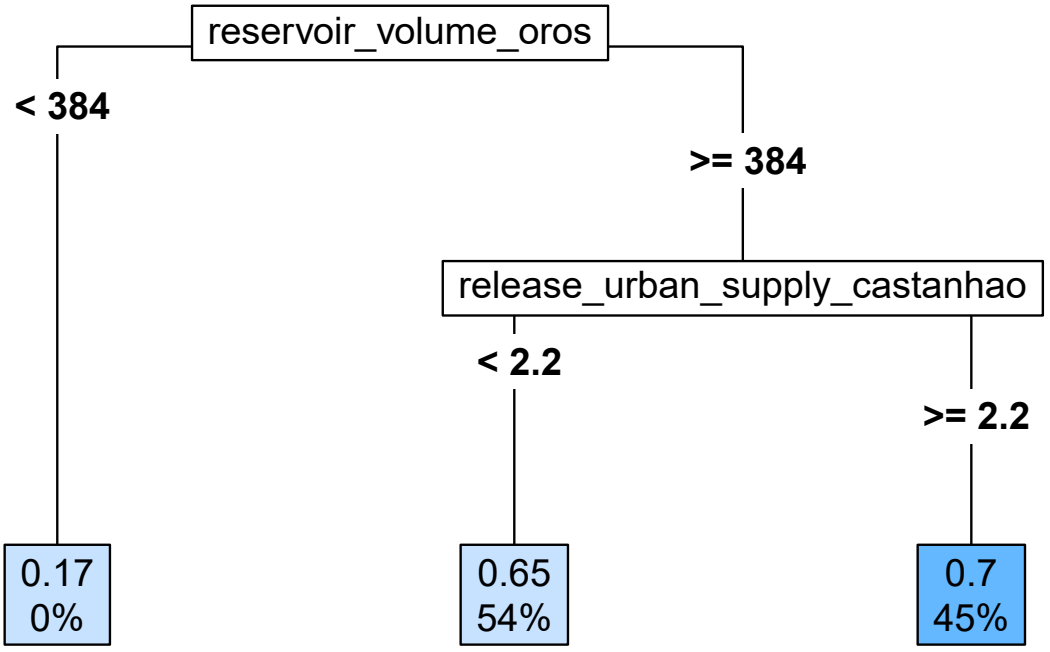
Response: Release for Urban supply − Castanhão
Month: November



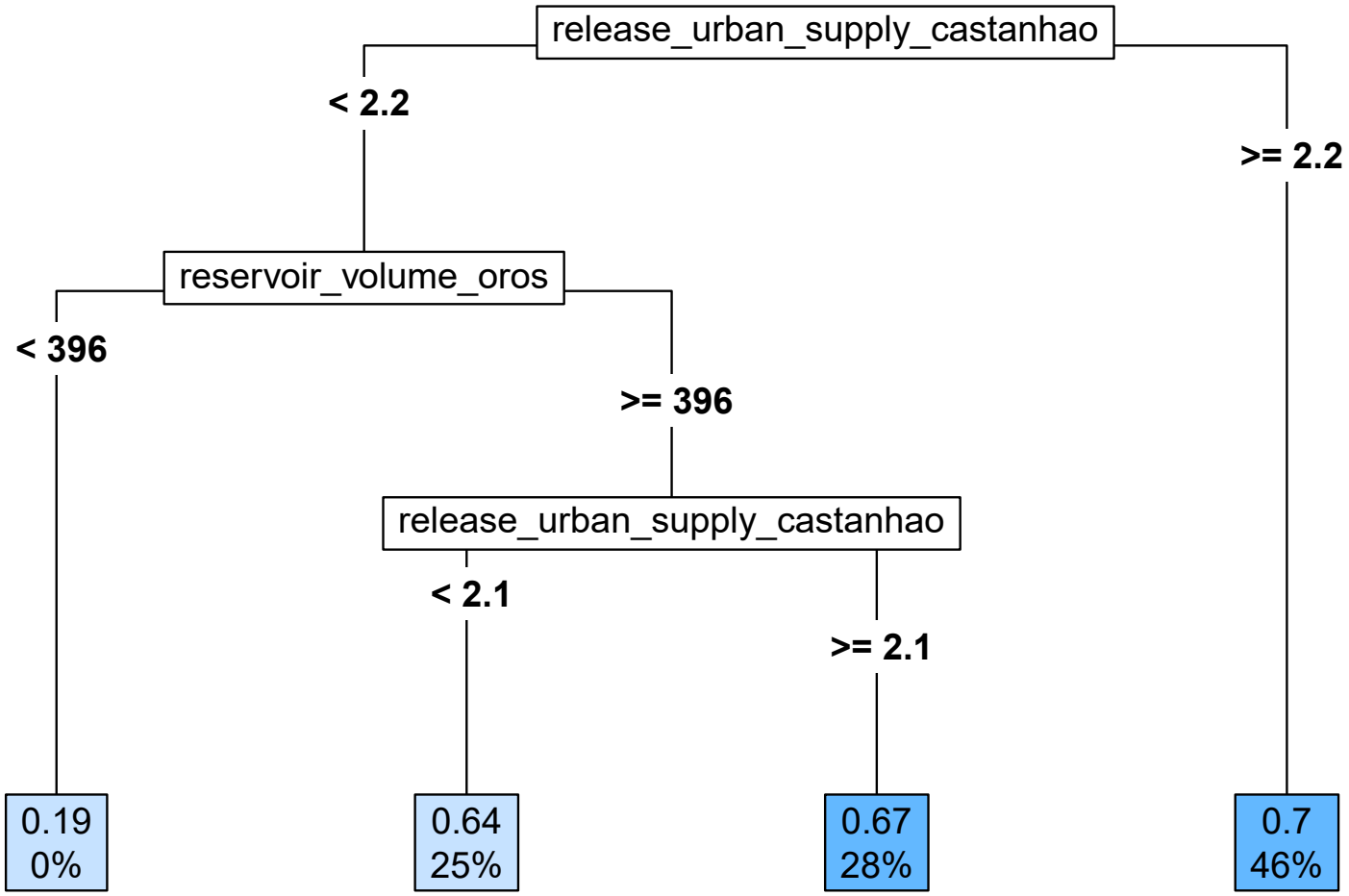
Response: Release for Urban supply − Castanhão
Month: December

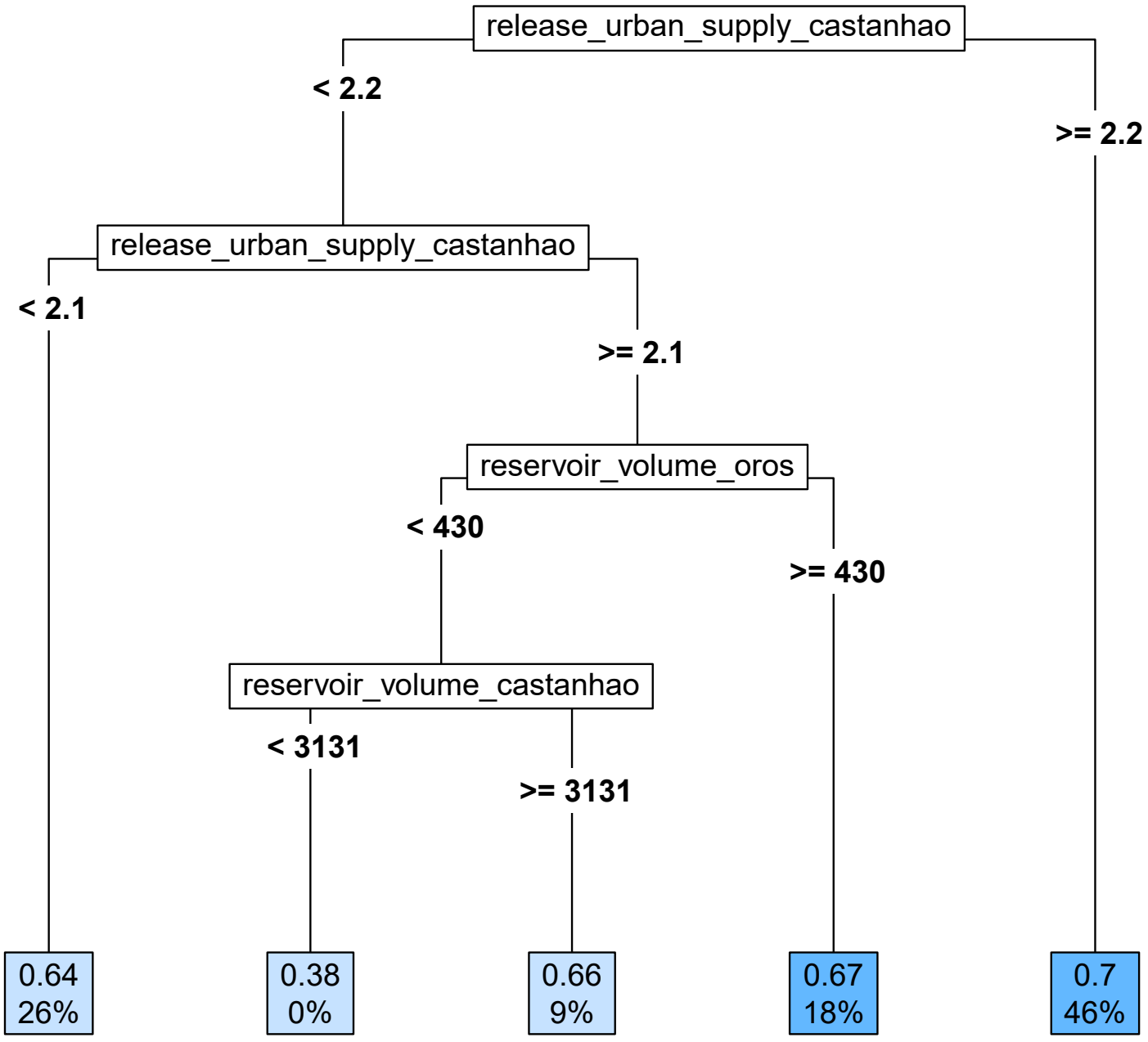


Response: Release for Urban supply − Orós
Month: January

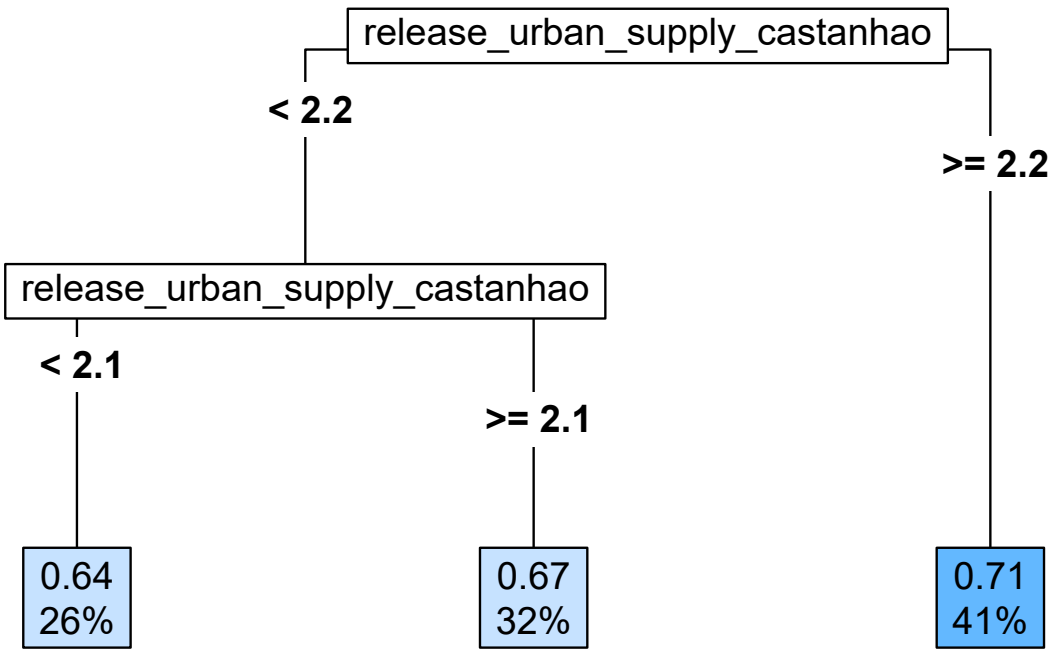


Response: Release for Urban supply − Orós
Month: February

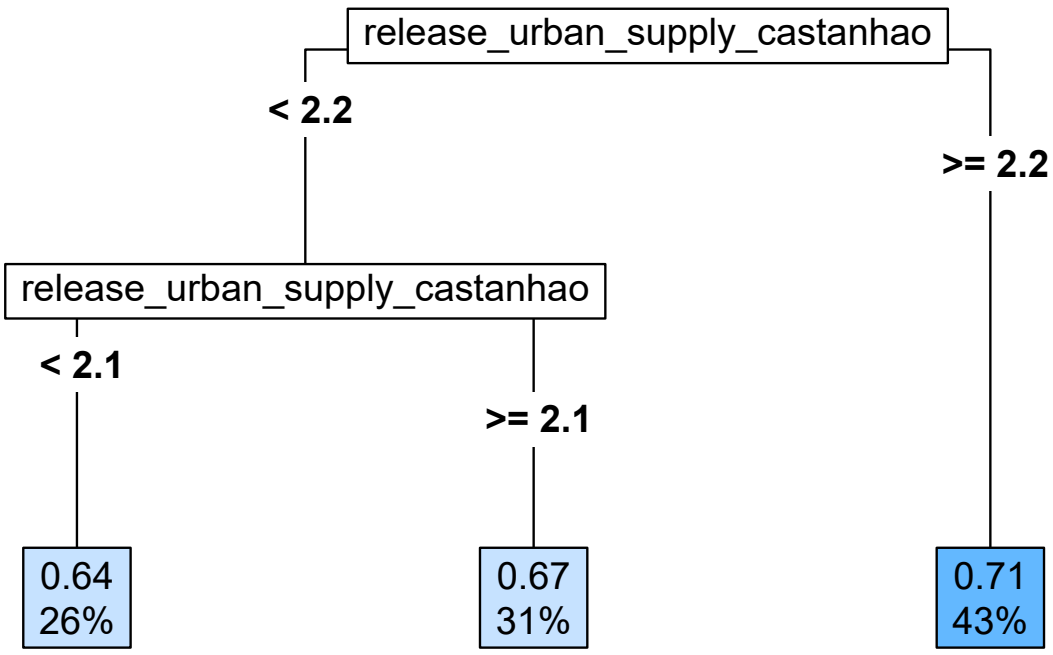



Response: Release for Urban supply − Orós
Month: March
release_urban_supply_castanhao
< 2.2
>= 2.2
release_urban_supply_castanhao
< 2.1
>= 2.1
reservoir_volume_oros
< 430
>= 430
reservoir_volume_castanhao
< 3131
>= 3131
0.64
26%
0.38
0%
0.66
9%
0.67
18%
0.7
46%

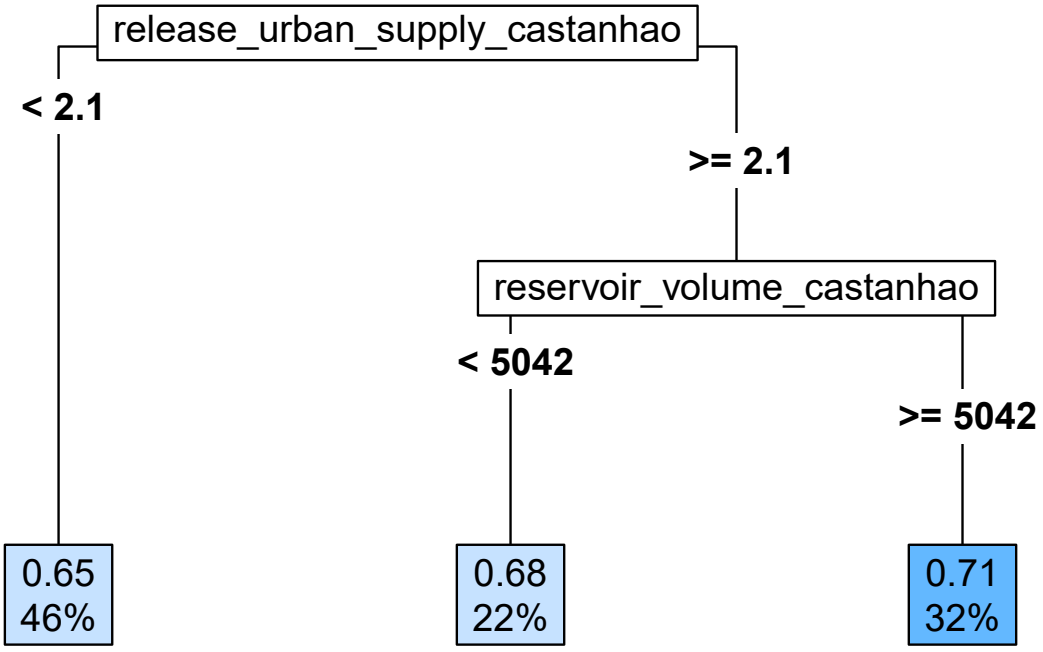
Response: Release for Urban supply − Orós
Month: April



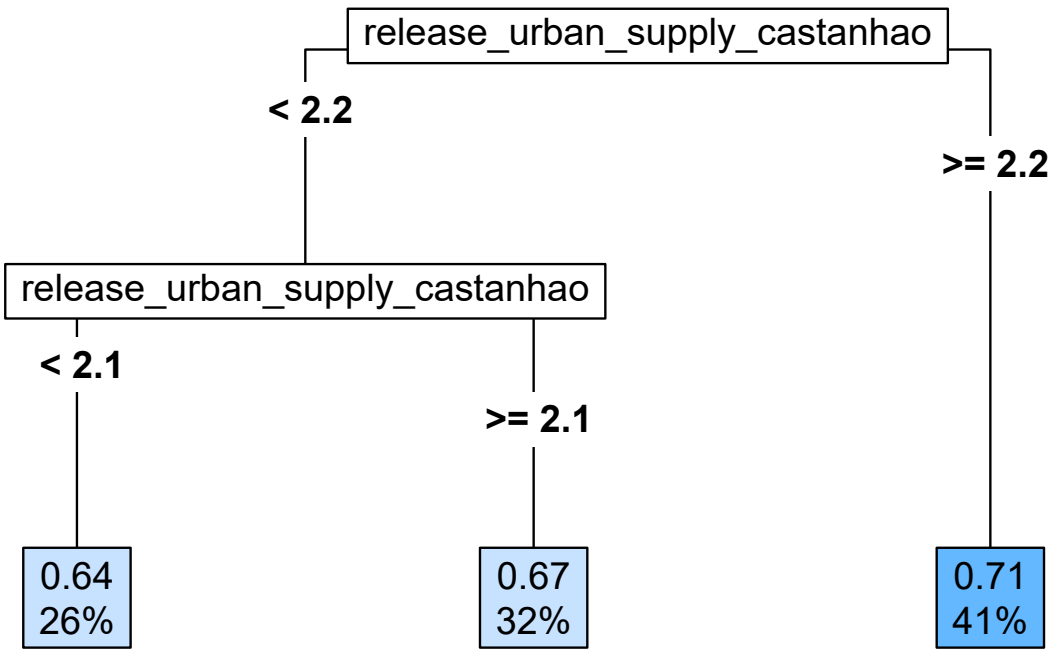
Response: Release for Urban supply − Orós
Month: May



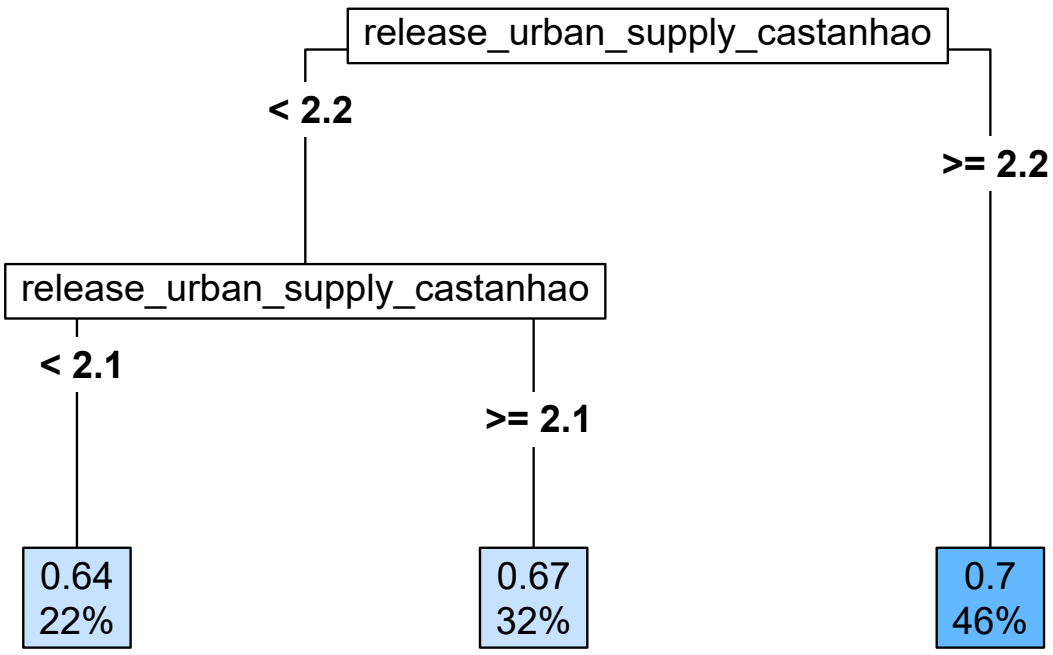
Response: Release for Urban supply − Orós
Month: June



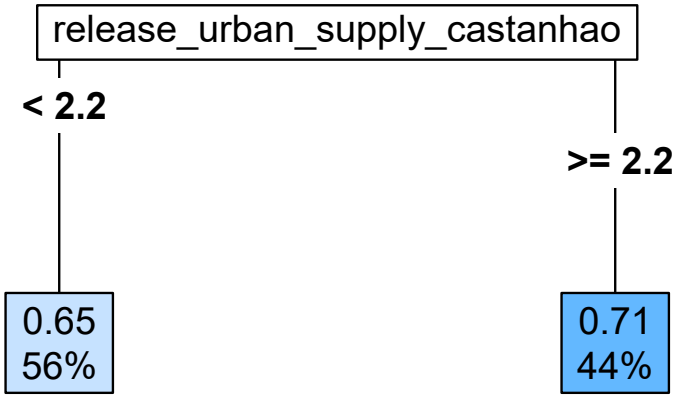
Response: Release for Urban supply − Orós
Month: July



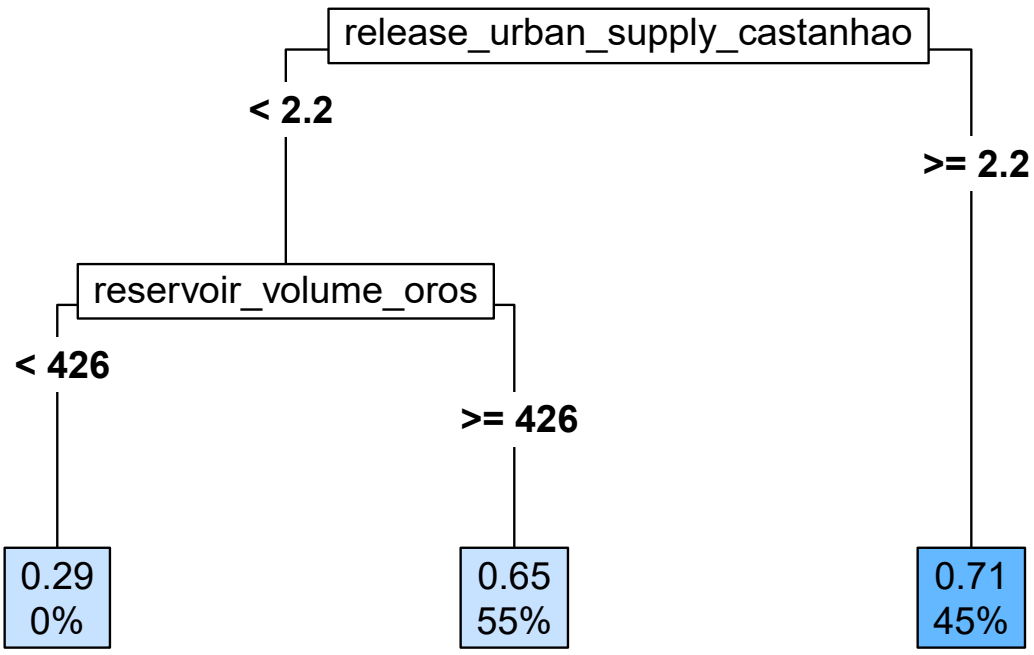
Response: Release for Urban supply − Orós
Month: August



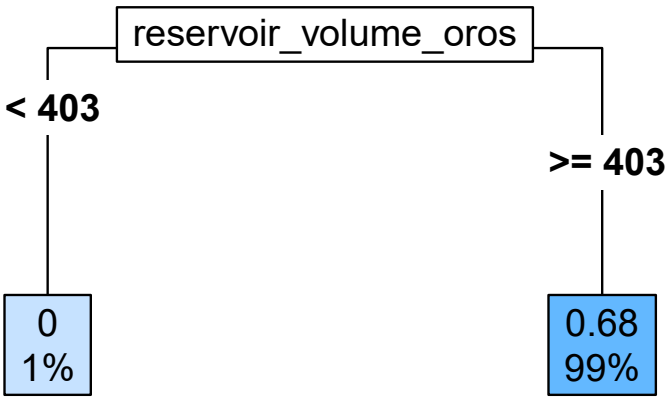
Response: Release for Urban supply − Orós
Month: September



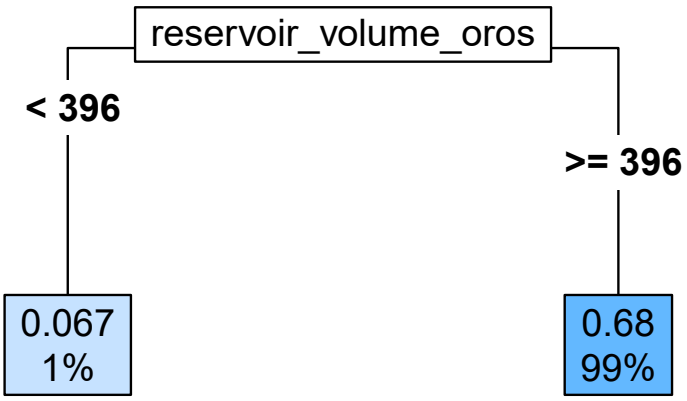
Response: Release for Urban supply − Orós
Month: October



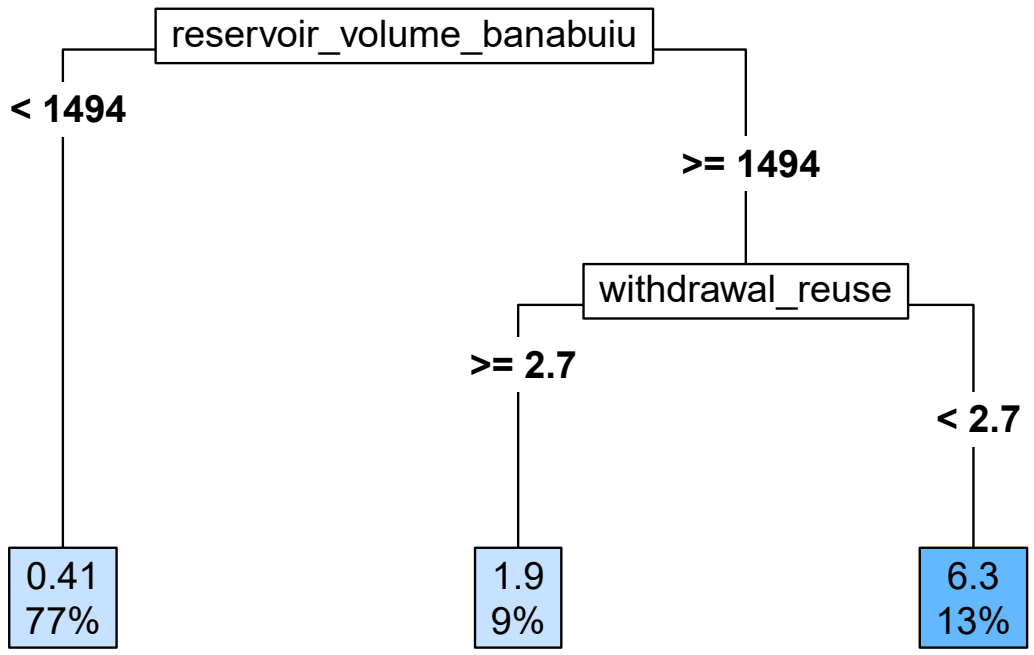
Response: Release for Urban supply − Orós
Month: November



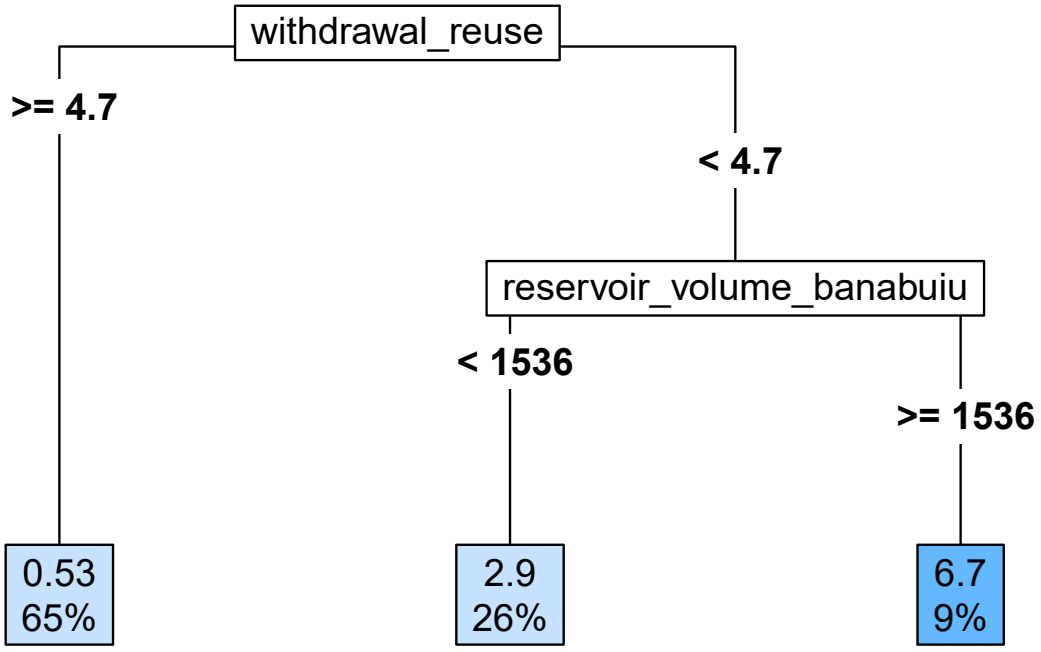
Response: Release for Urban supply − Orós
Month: December



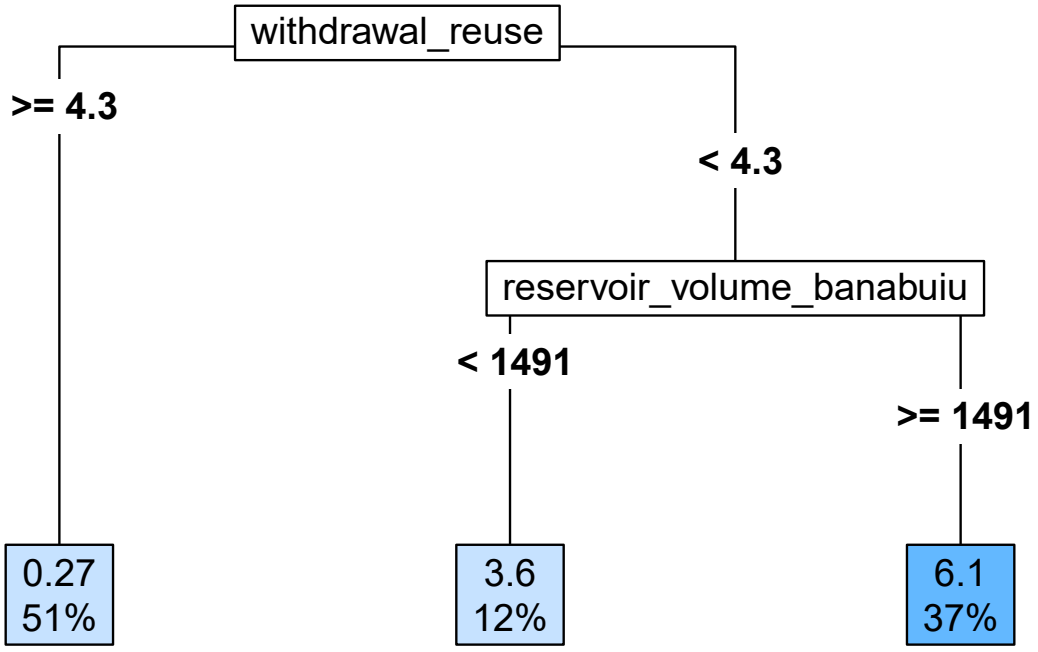
Response: Release for Urban supply − Sítios Novos
Month: January



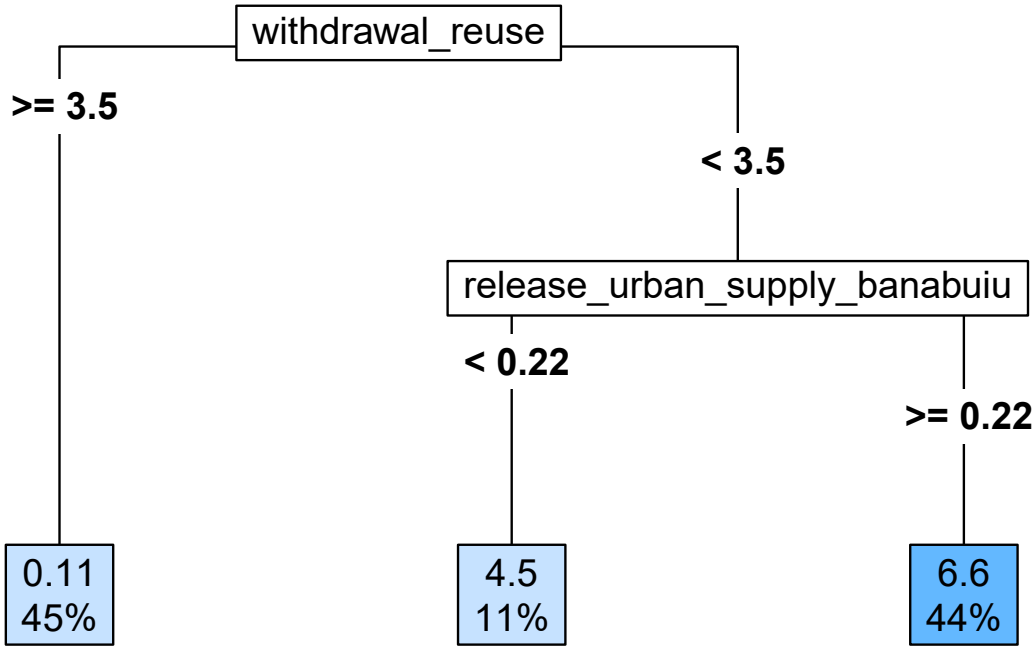
Response: Release for Urban supply − Sítios Novos
Month: February



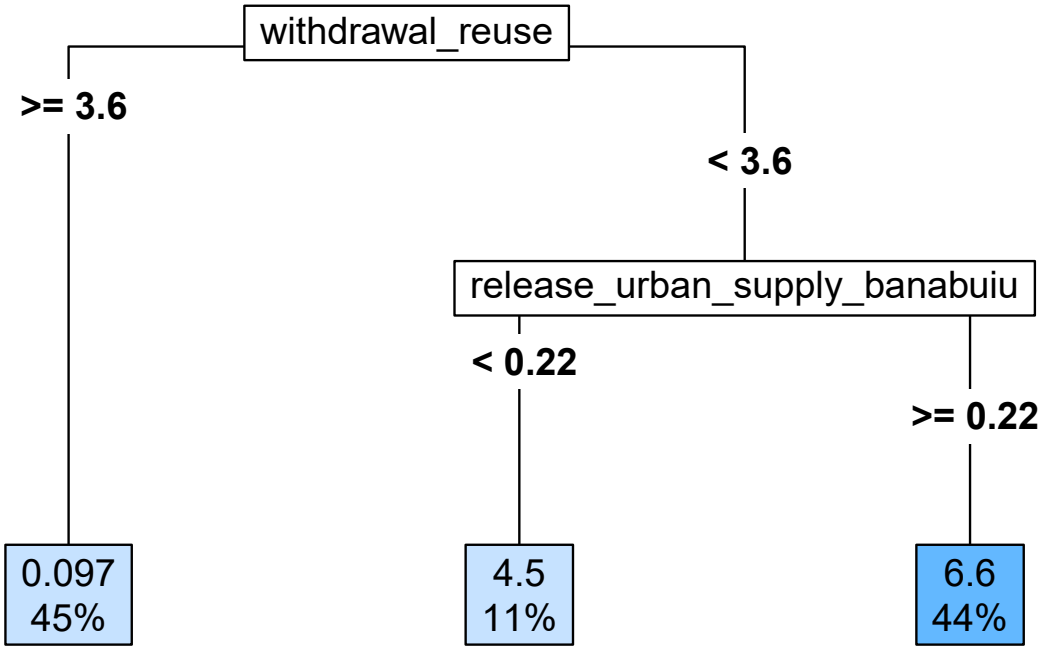
Response: Release for Urban supply − Sítios Novos
Month: March



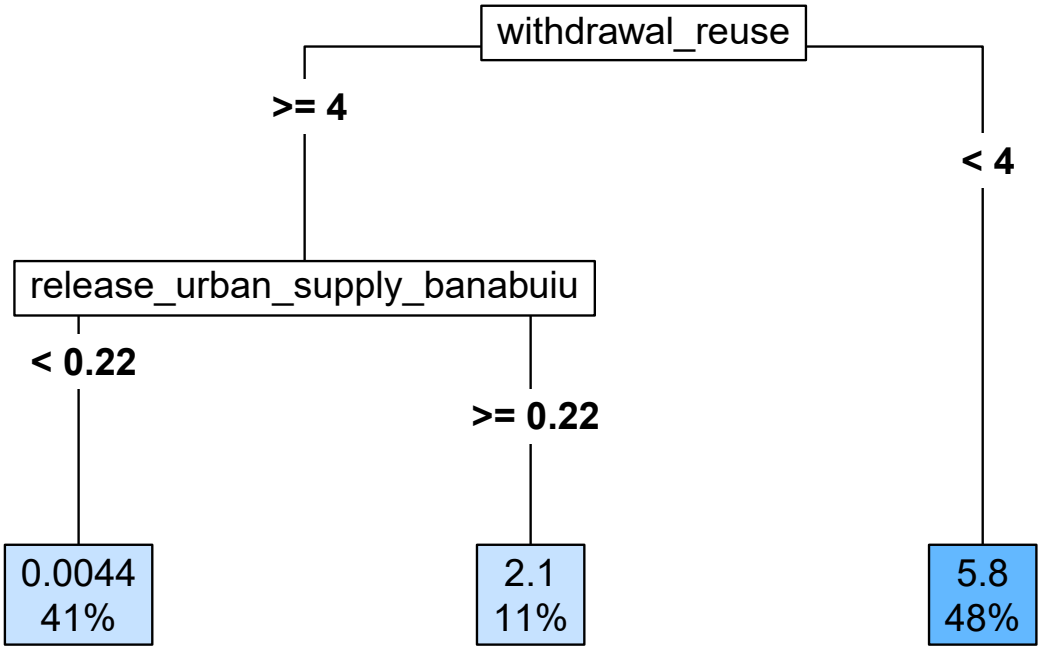
Response: Release for Urban supply − Sítios Novos
Month: April



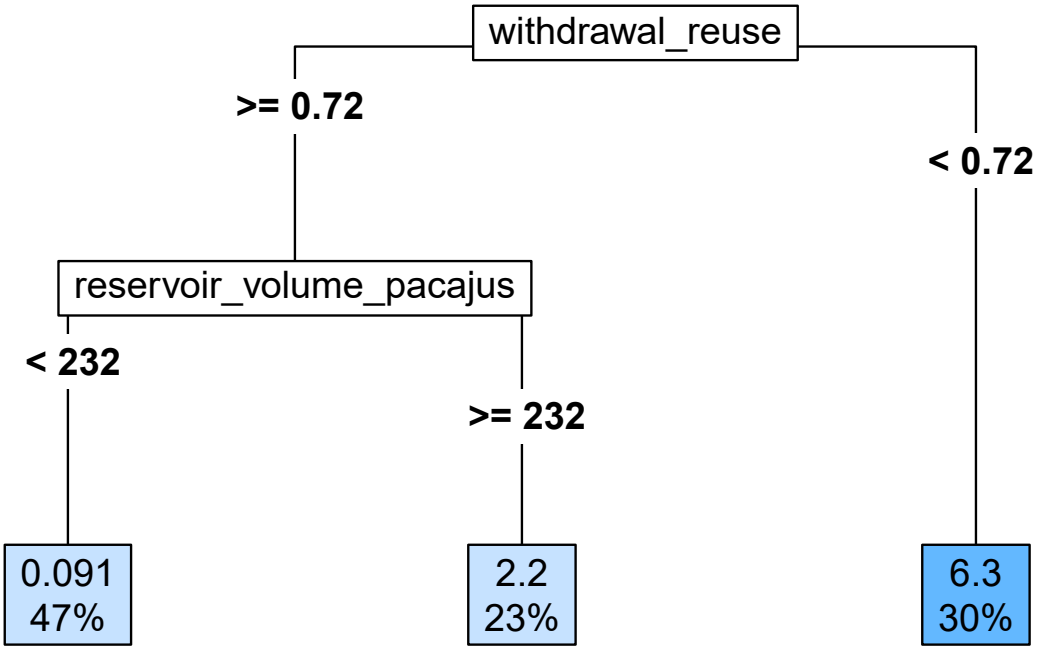
Response: Release for Urban supply − Sítios Novos
Month: May



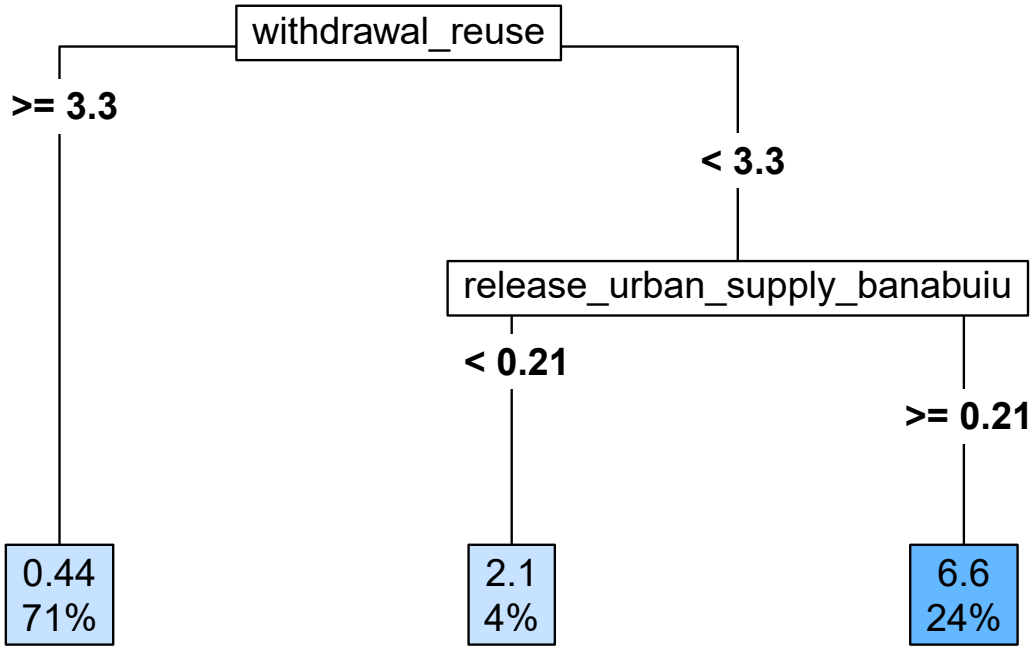
Response: Release for Urban supply − Sítios Novos
Month: June



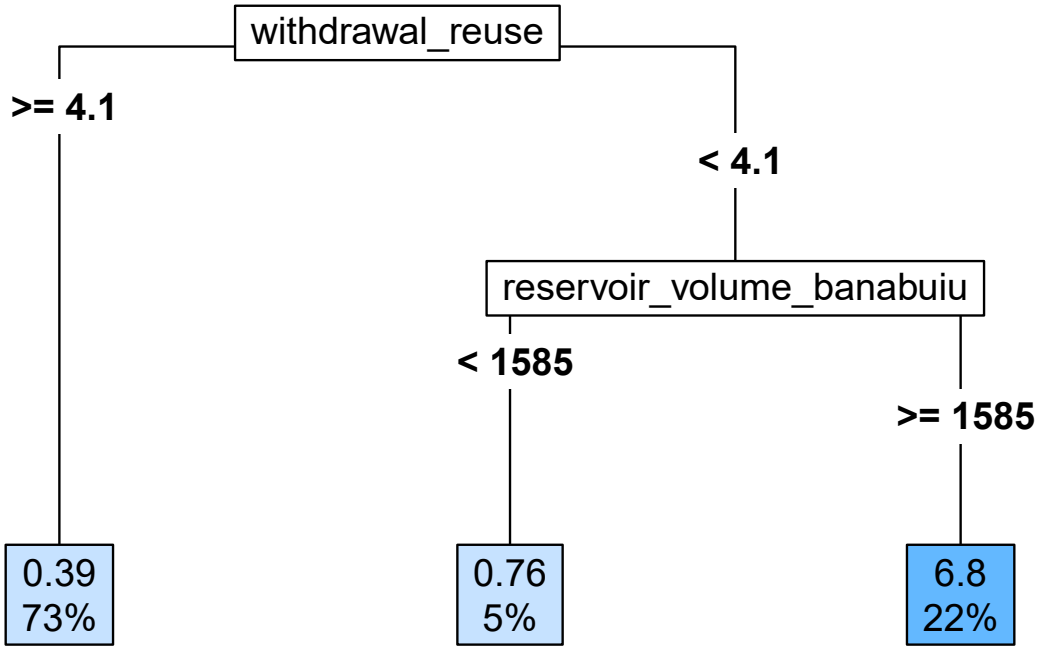
Response: Release for Urban supply − Sítios Novos
Month: July



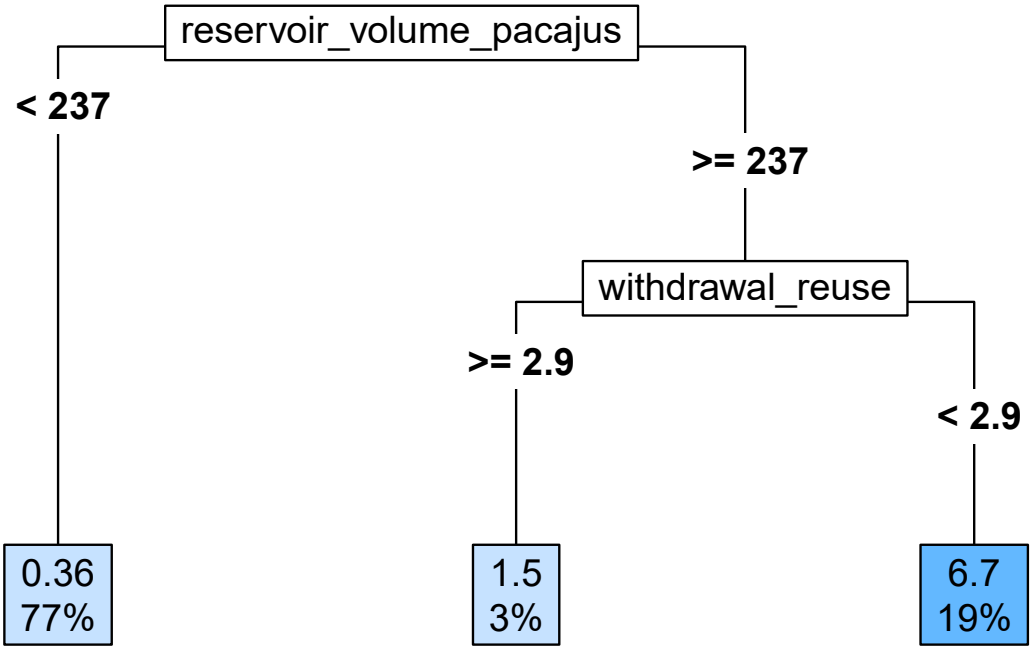
Response: Release for Urban supply − Sítios Novos
Month: August



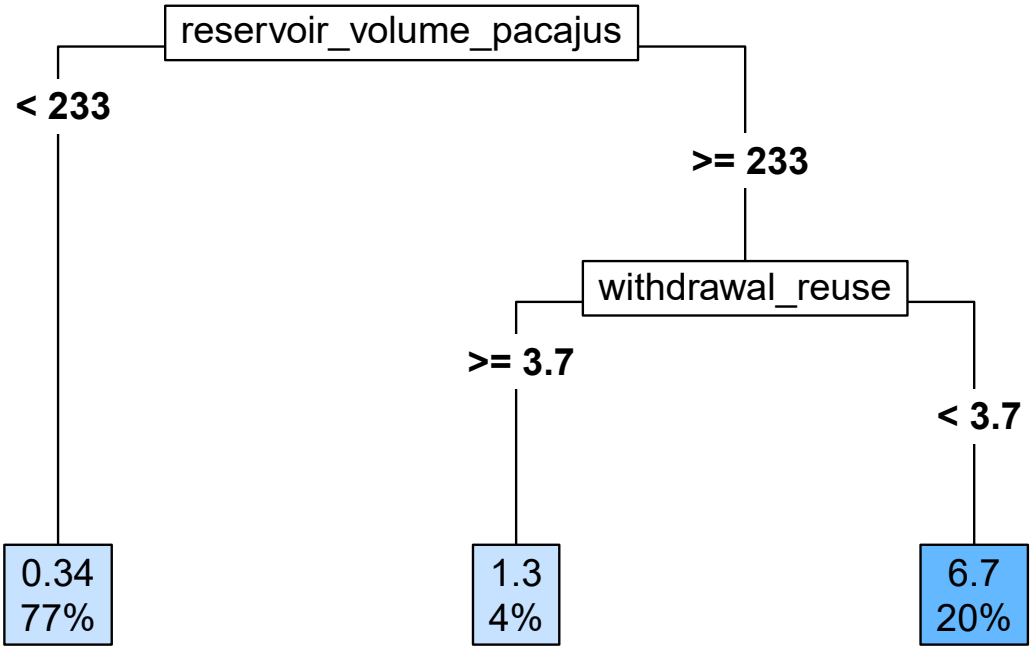
Response: Release for Urban supply − Sítios Novos
Month: September



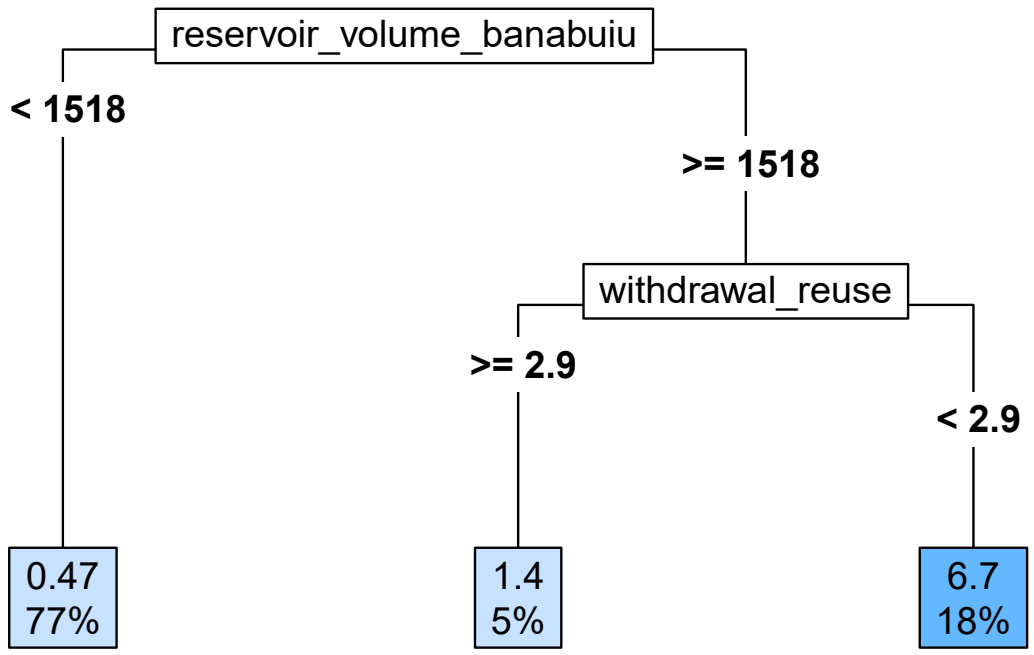
Response: Release for Urban supply − Sítios Novos
Month: October



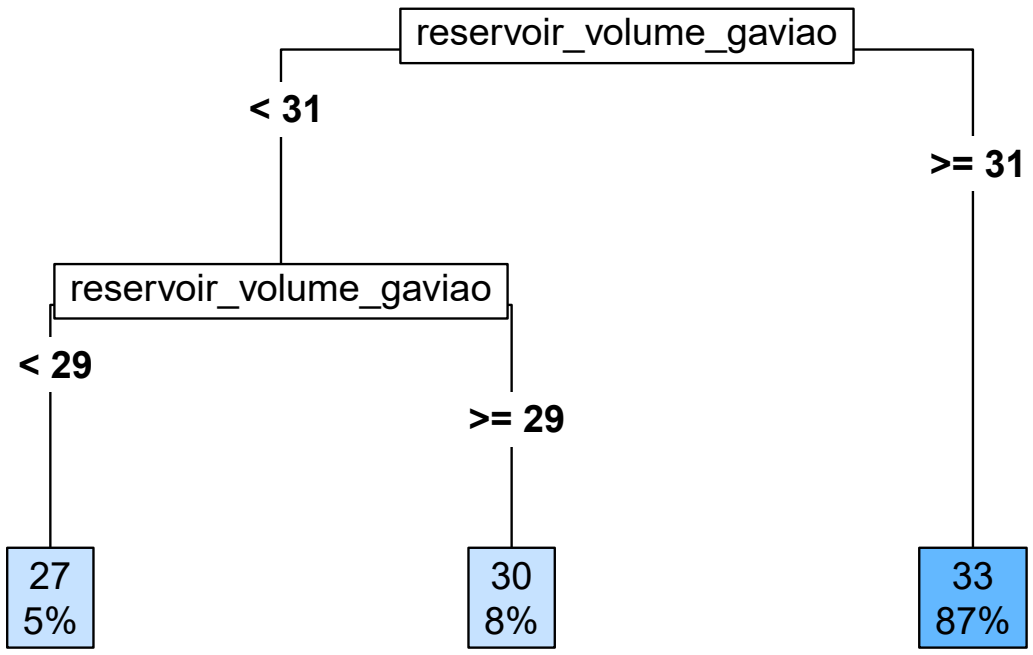
Response: Release for Urban supply − Sítios Novos
Month: November



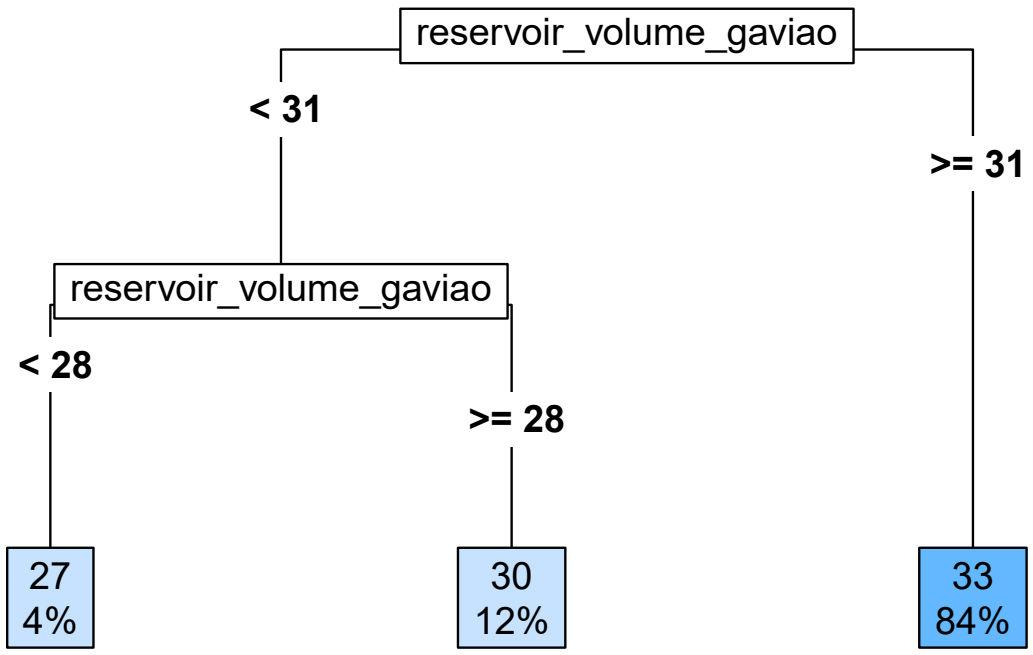
Response: Release for Urban supply − Sítios Novos
Month: December



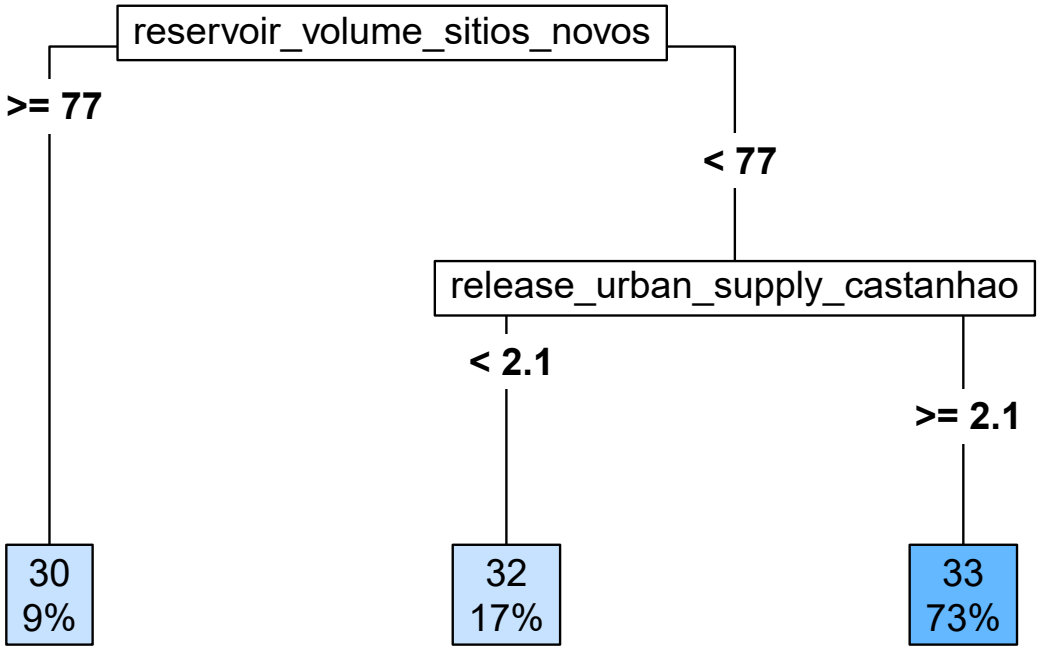
Response: Release for Urban supply − Gavião
Month: January



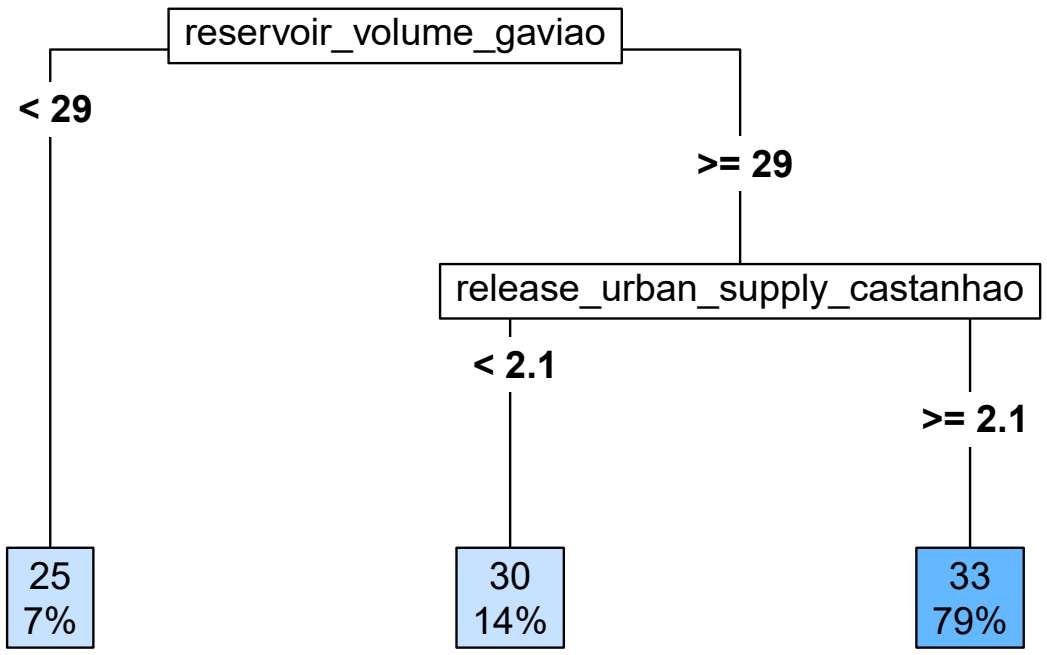
Response: Release for Urban supply − Gavião
Month: February



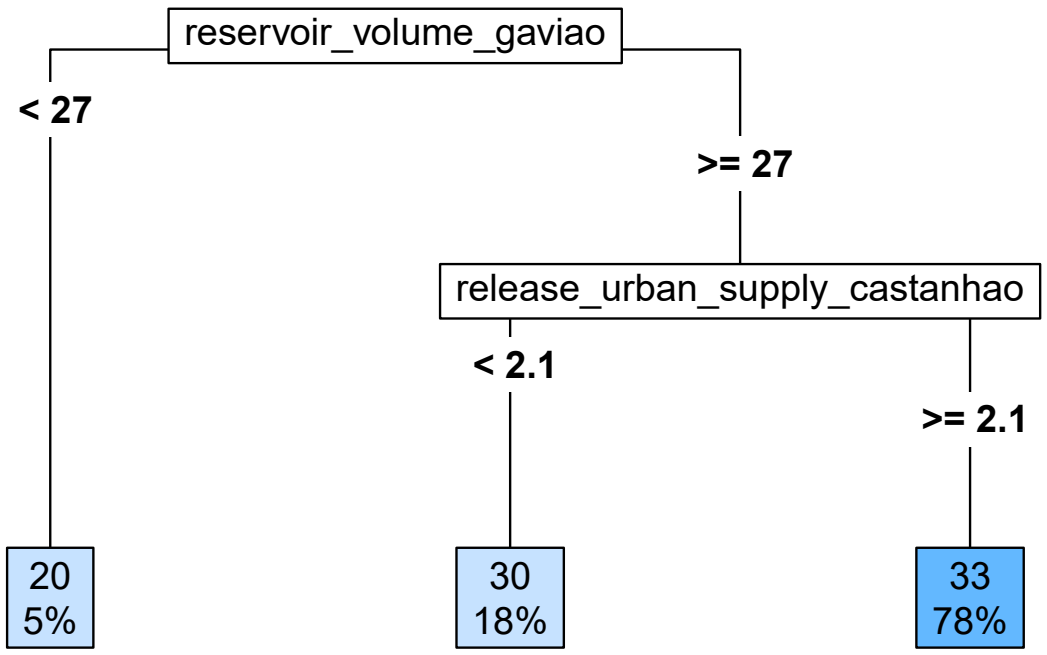
Response: Release for Urban supply − Gavião
Month: March



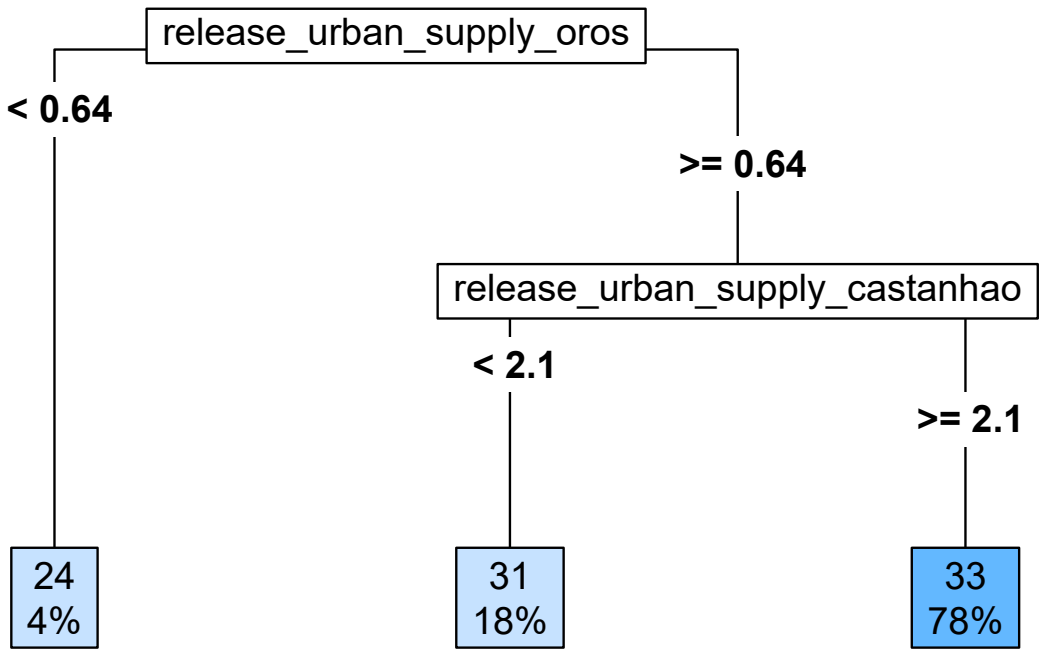
Response: Release for Urban supply − Gavião
Month: April



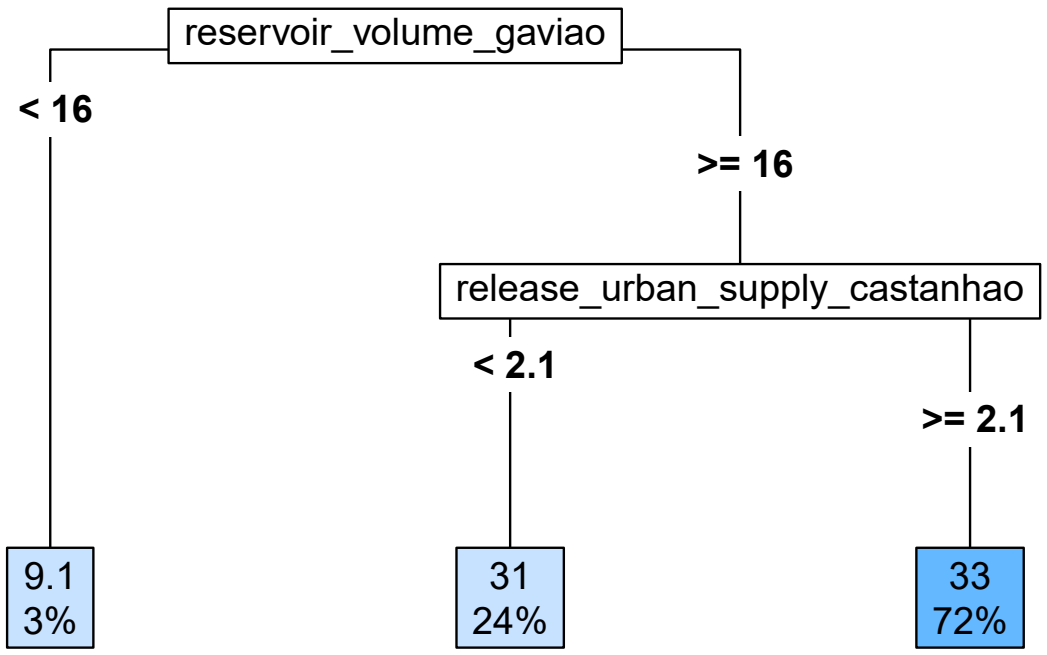
Response: Release for Urban supply − Gavião
Month: May



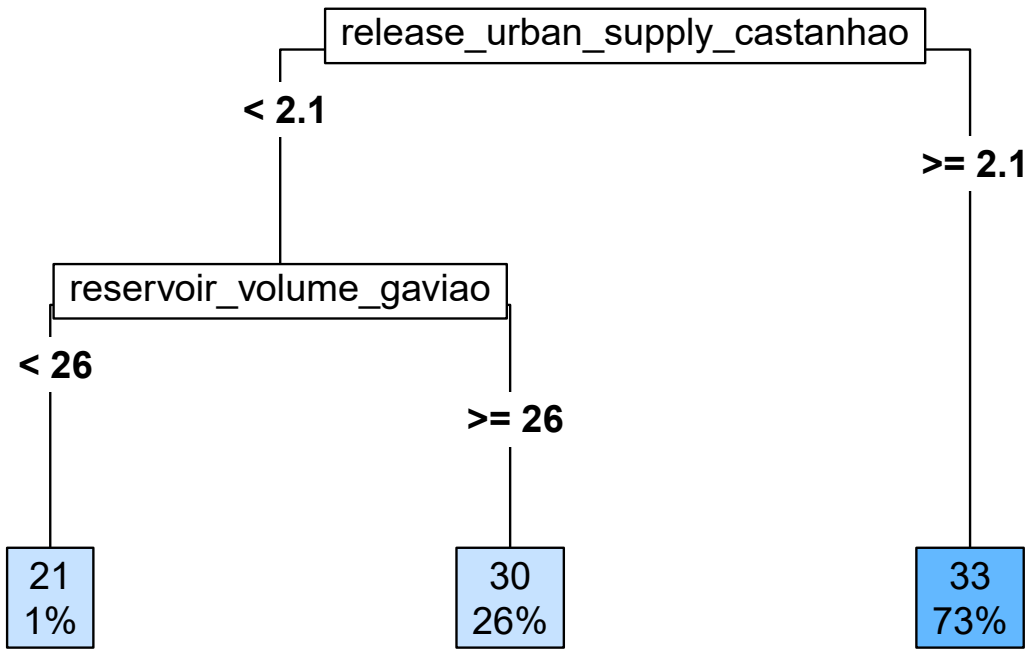
Response: Release for Urban supply − Gavião
Month: June



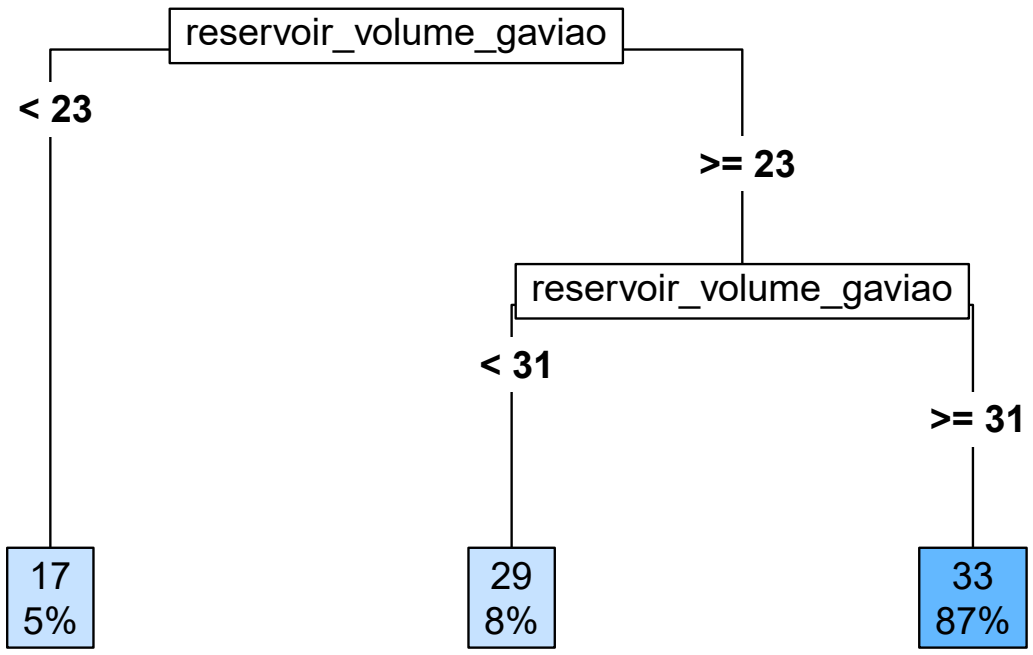
Response: Release for Urban supply − Gavião
Month: July



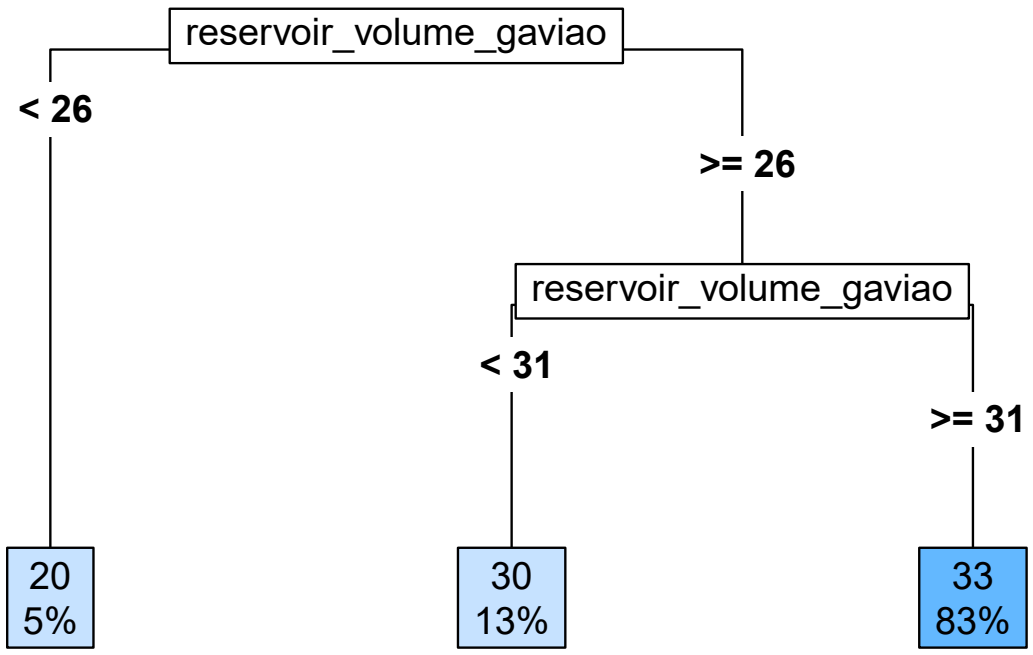
Response: Release for Urban supply − Gavião
Month: August



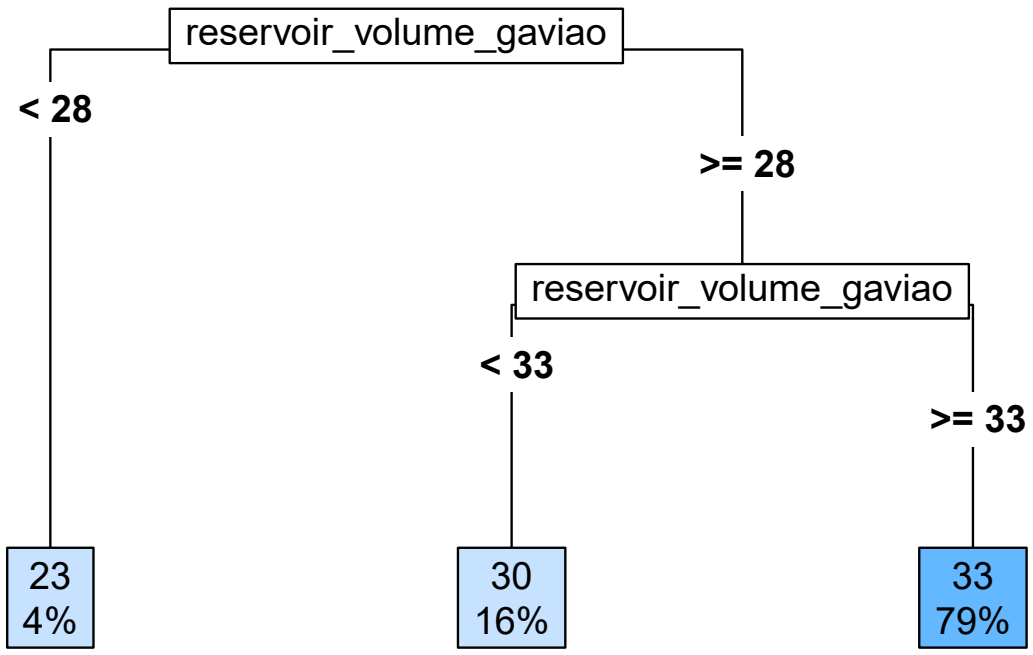
Response: Release for Urban supply − Gavião
Month: September



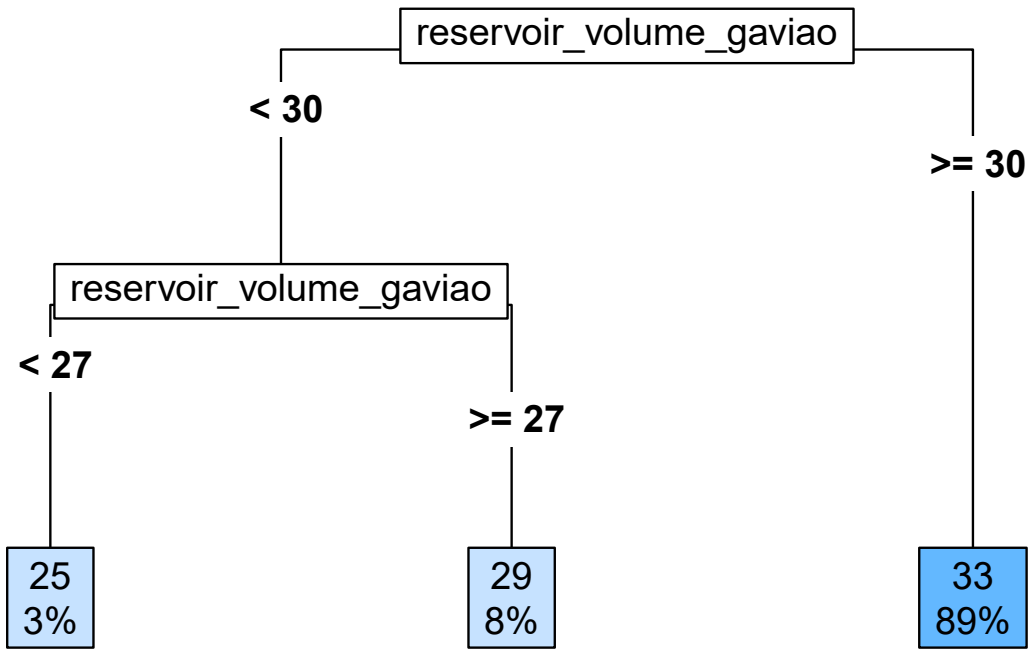
Response: Release for Urban supply − Gavião
Month: October



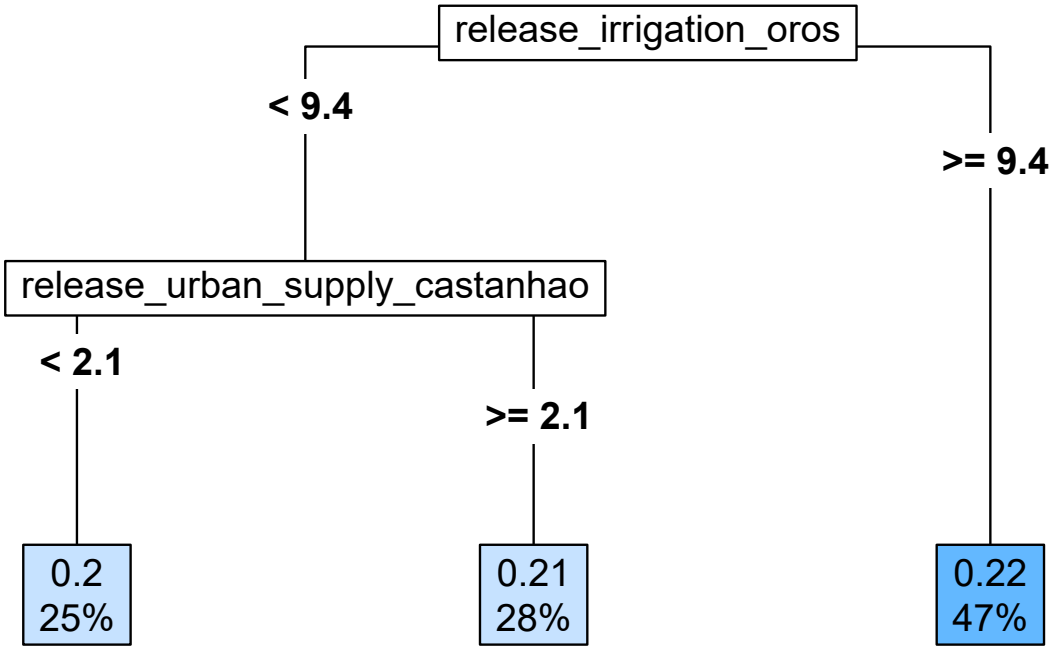
Response: Release for Urban supply − Gavião
Month: November



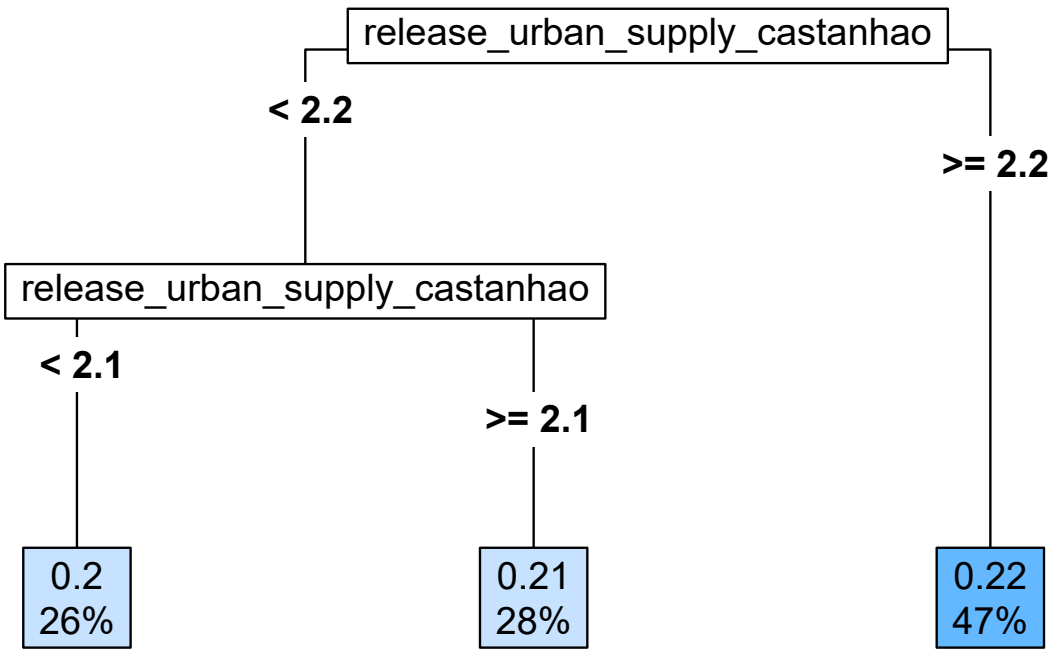
Response: Release for Urban supply − Gavião
Month: December



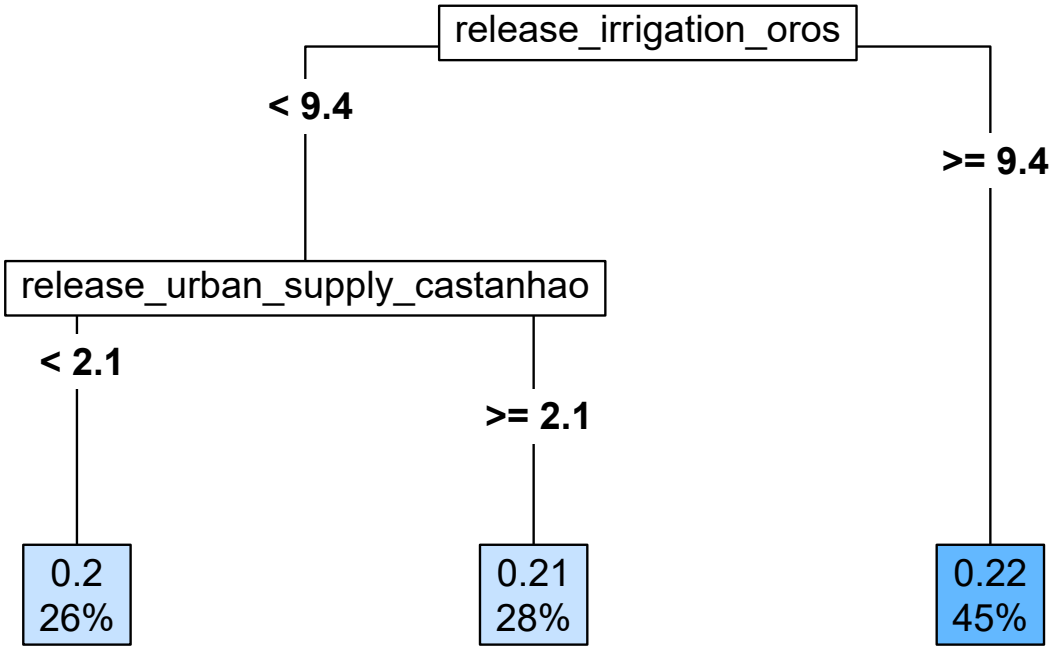
Response: Release for Urban supply − Banabuiú
Month: January



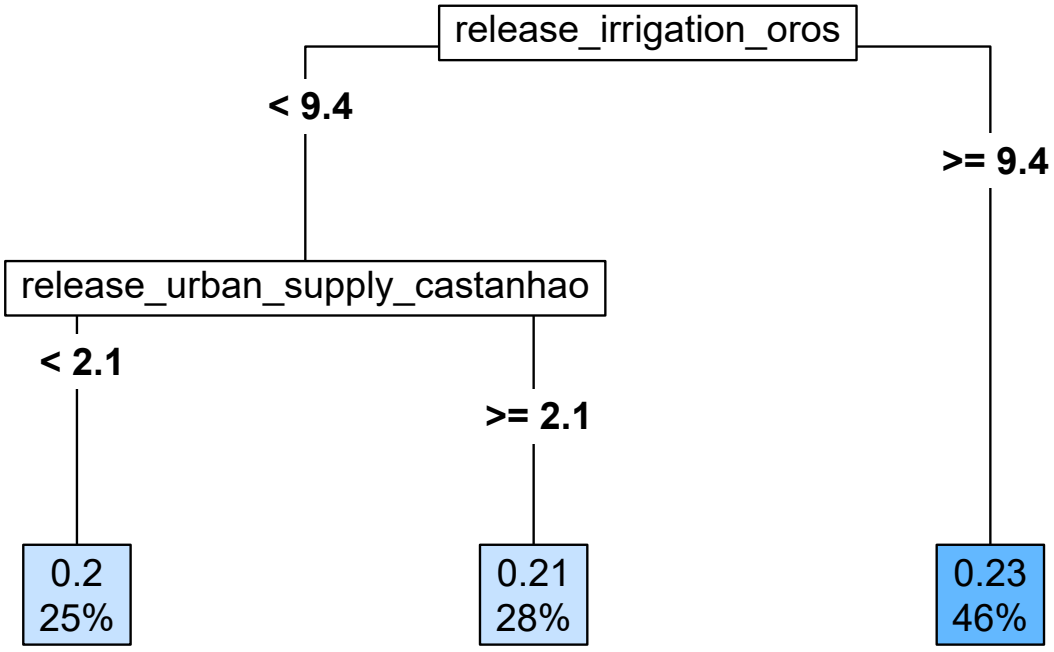
Response: Release for Urban supply − Banabuiú
Month: February



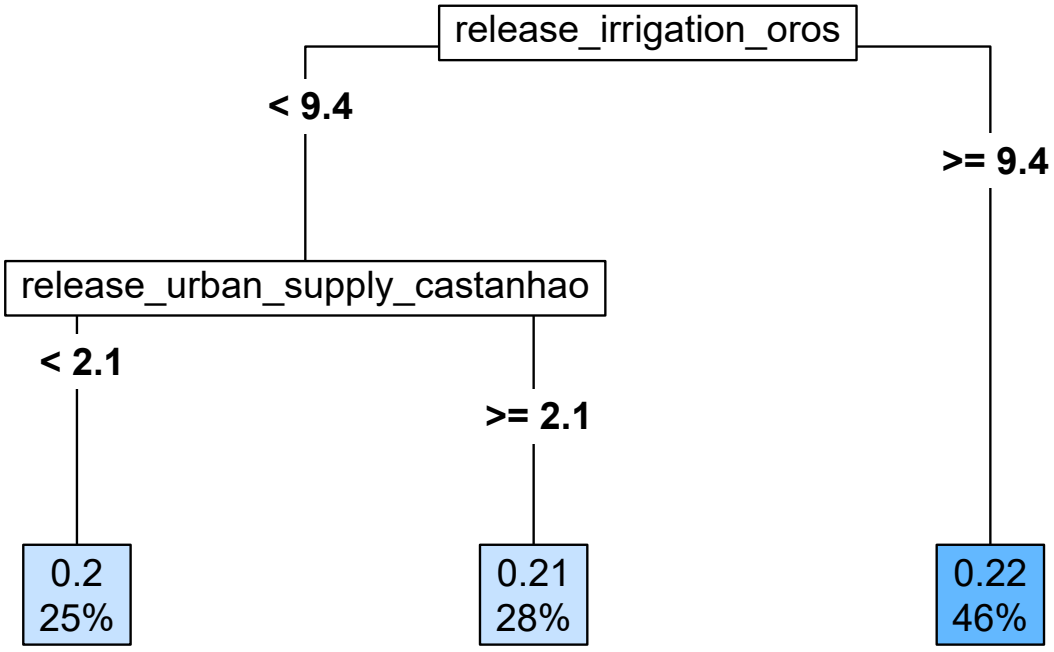
Response: Release for Urban supply − Banabuiú
Month: March



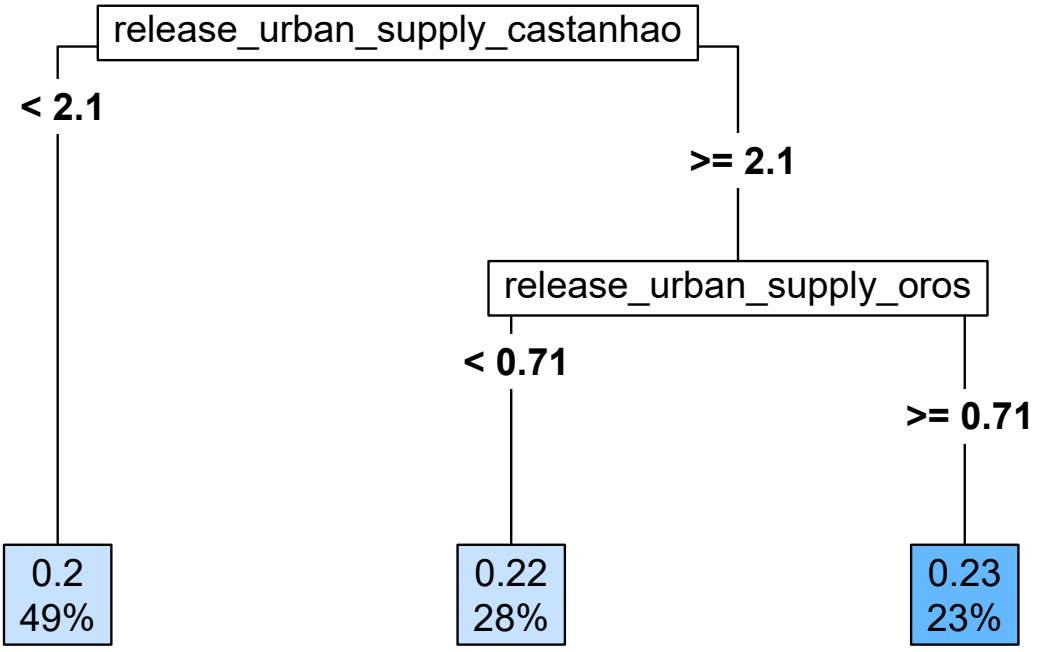
Response: Release for Urban supply − Banabuiú
Month: April



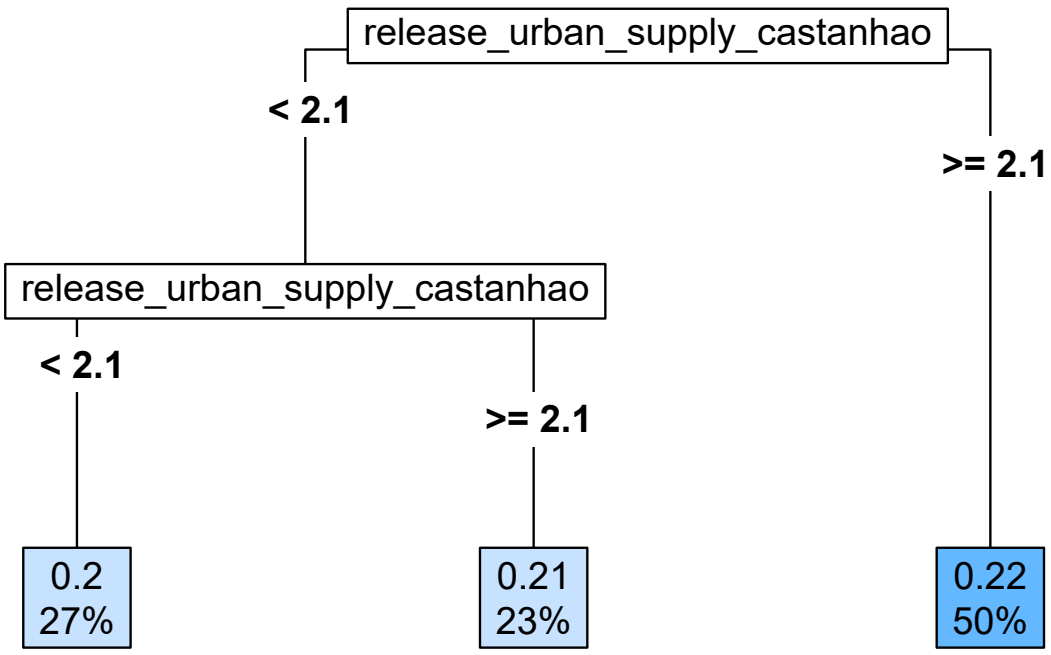
Response: Release for Urban supply − Banabuiú
Month: May



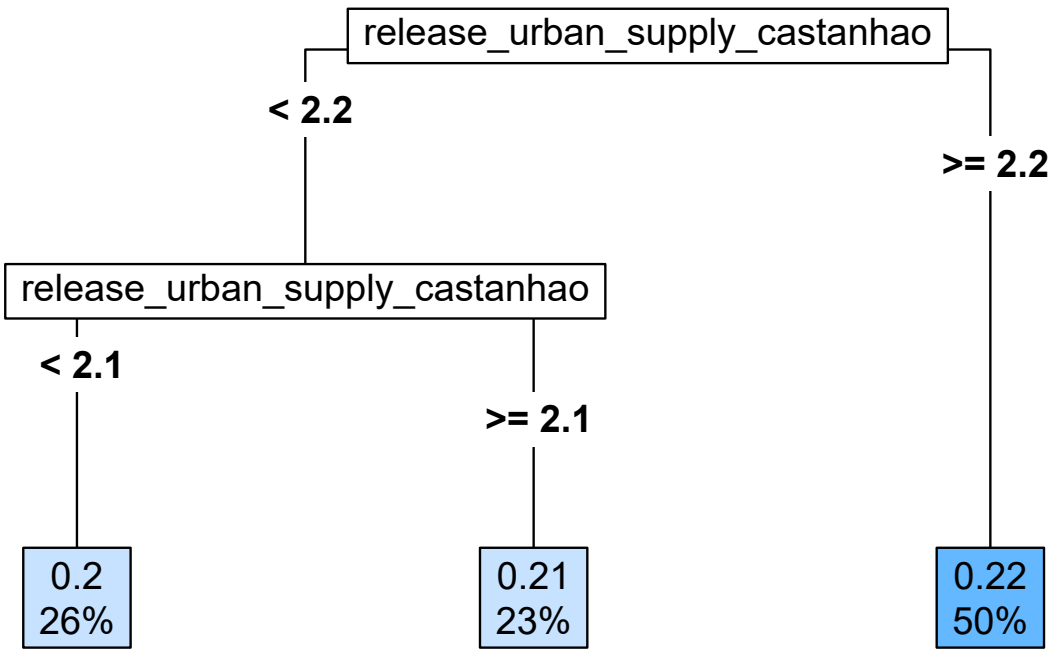
Response: Release for Urban supply − Banabuiú
Month: June



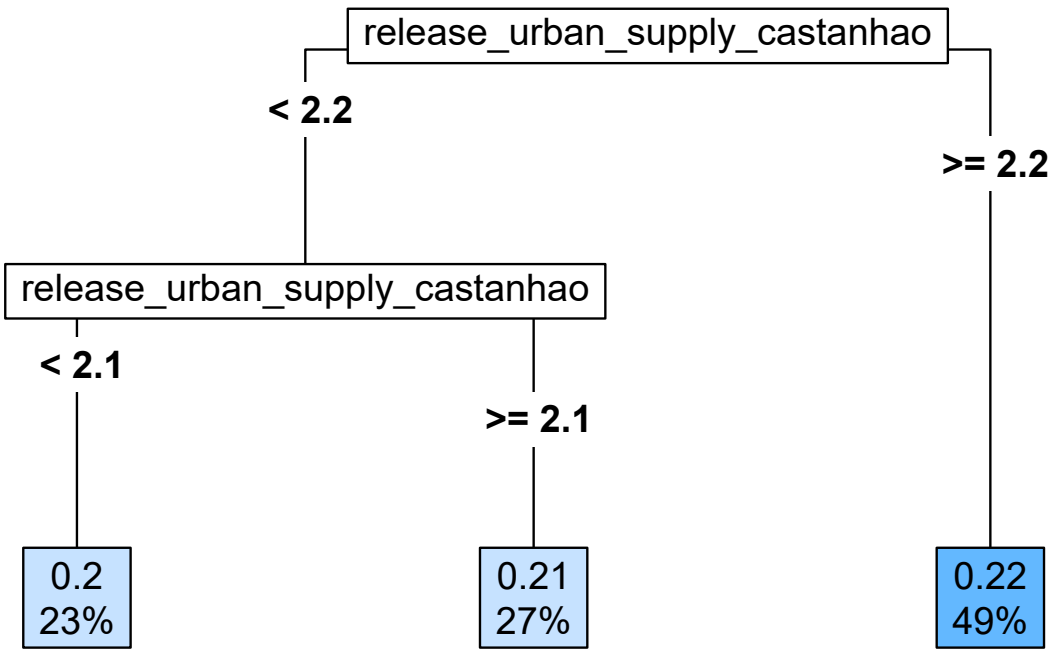
Response: Release for Urban supply − Banabuiú
Month: July



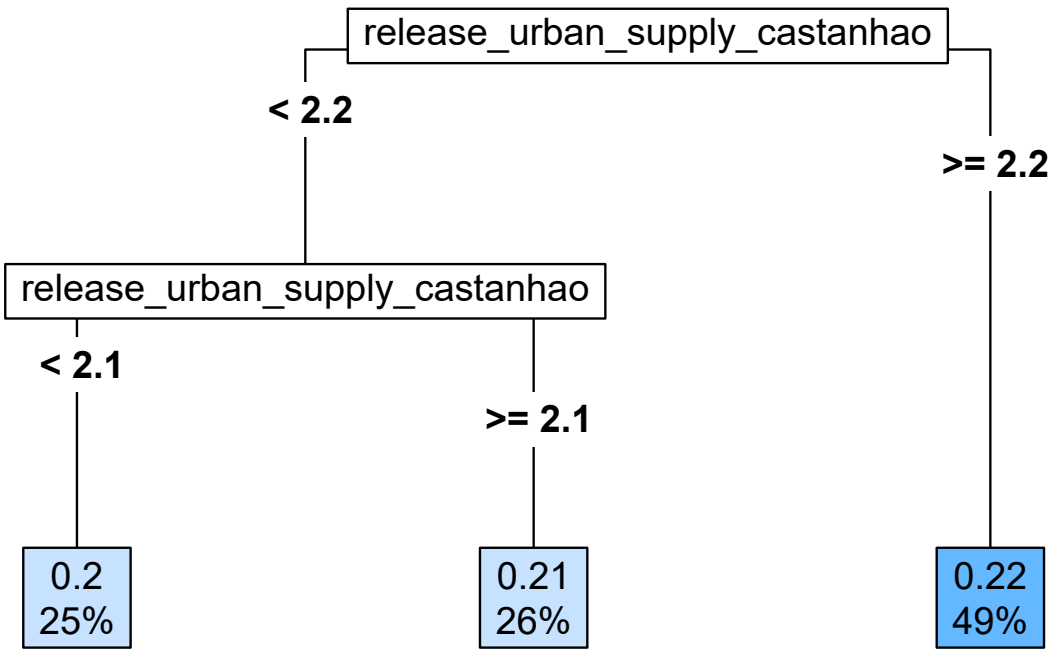
Response: Release for Urban supply − Banabuiú
Month: August



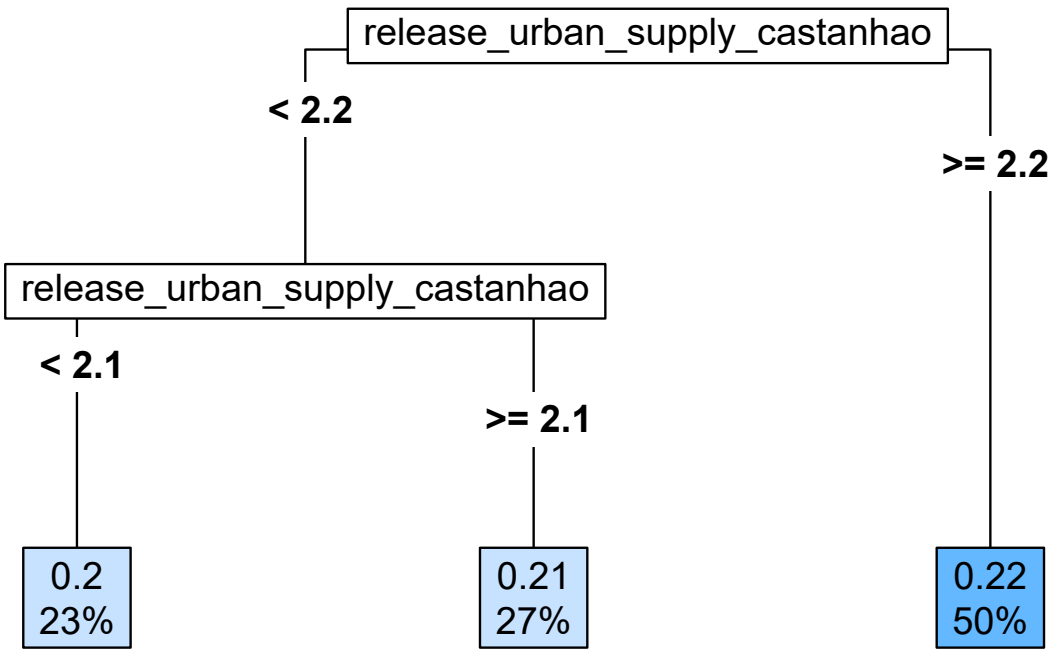
Response: Release for Urban supply − Banabuiú
Month: September



Response: Release for Urban supply − Banabuiú
Month: October



Response: Release for Urban supply − Banabuiú
Month: November



Response: Release for Urban supply − Banabuiú
Month: December