Urban Verticalization and Water Consumption: A Data-Driven approach for São Paulo

Arthur Hiratsuka Rezende[1] (✉) and André P. de L. F. de Carvalho[1]

Institute of Mathematical and Computer Sciences, University of São Paulo, São Paulo, Brazil arthurhr@usp.br, andre@icmc.usp.br

Abstract. Concurrent trends of urbanization and population growth in Brazil can exert high pressure on the (already degraded) environment. In the city of São Paulo, in particular, there is a clear trend towards verticalization of real estate, increasing population density. To attend demands due to this rapid change in a particular area, water consumption, it is necessary to understand the aspects related with domestic water demand. The main objective of this study is to analyze the monthly water consumption in high-rise residential properties, and investigate the descriptive power of building related variables using machine learning. For such, real consumption data from the past three years (provided by the water and sewage company Sabesp) were obtained, along with two databases containing detailed information on high-rise apartment buildings in the city of São Paulo. After a meticulous integration of these databases, reliable information were obtained for 3,299 high-rise buildings, totalizing 276,670 apartments, described by 21 variables. One potential weakness in commonly used estimates (e.g., demographic, financial) is that they may be outdated or biased. In contrast, the physical characteristics of buildings are easily verifiable, and simple to obtain. The study's hypothesis is that relying solely on the building features may preserve a similar descriptive power, while eliminating uncertainties and biases. A contribution of this study is the estimation of the monthly consumption per unit, which can be used for modeling urban water distribution systems. In the experiments carried out, fourteen different regression algorithms for consumption prediction are investigated, and the predictive performance of the induced models is comparable with similar studies that use building characteristics alongside population estimates and water/sewage features in the building, partially confirming the research hypothesis.

Keywords: Water Consumption · Urban Infrastructure · Urban Verticalization · Machine Learning

1 Introduction

The rational management of natural resources is crucial for humanity. The United Nations[1] estimates that the global population reached 8 billion in 2022,

[1] UN-Habitat report - 2022.

with projections of nearly 10 billion by 2050 25% increase in just 28 years. According to the report, approximately 58.3% of the global population, including 54.3% in less developed regions, is expected to live in cities by 2025. These proportions are projected to rise to 68.4% and 65.6%, respectively, by 2050.

Additionally, according to the 2023 SNIS report[2], national water losses reached 37.8%, with São Paulo reporting a loss rate of 27.9%[3]. In this context, accurate consumption estimates are essential for realistic system modeling and simulation. These tools can support loss reduction efforts through improved leak detection.

Regarding water distribution (WD) in Brazil, demographic trends significantly influence water consumption patterns, as demonstrated in the Section 4.1. This study is part of a preliminary investigation supporting the revision of the Technical Standard for Residential Building Design by Sabesp, the water and sewage company that serves the city of São Paulo, in São Paulo state, Brazil. The study includes a diagnostic assessment of the current predictors performance and presents preliminary results on the development of a new regressor.

1.1 Urbanization and its Effects on Water Consumption

The relationship between water consumption and urban density has been verified in the city of Barcelona, Spain [6], due to population shifts from the central to peripheral areas. Urban areas in the city of Hawassa, Ethiopia [13], increased from 7.2% of the territory in 1991 to 26.5% in 2021, estimated to reach 45.9% by 2051. This expansion is expected to drive a 20% increase in water consumption.

A simulation of land-use changes in Brazil [3] projects an urban area expansion of over 4 km^2 in the city of Campina Grande, Brazil (an 8% increase) at the expense of rural areas. This transformation is expected to lead to a 7% rise in water consumption in the city between 2020 and 2050. In Tehran, Iran [30], deteriorated areascharacterized by structurally unstable buildings and streets with limited accessibilityshow a negative correlation with water consumption, which may be attributed to outdated and inefficient distribution infrastructure.

Another factor that can affect urban WD, without increasing built-up areas, is the retrofitting of abandoned buildings into affordable housing, as studied by [7] in São Paulo. The city also has a housing policy of creating apartments of up to 50m^2 for social housing, and its 2014 Directive Plan promotes densification near public transport, driving the population to live closer to their working places, which is highly concentrated around the city center.

Since 1929, the city of São Paulo has seen a fast increasing number of high-rise apartment buildings (verticalization) [2] which accelerated during the 1960s and 1970s, and, since the 2000s, has seen an average of more than 250 new high-rise apartment buildings launched per year. By looking at the relationship between affordability and urban verticalization in the city, [18], it also be observed a paradigm shift in the recent constructions: a transition from low-to-mid-rise buildings to mid-to-high-rise developments, influencing both property prices and housing affordability.

[2] Sistema Nacional de Informações sobre Saneamento
[3] Sabesp report on water losses

1.2 Main Contributions

The main contribution from this study is to characterize the consumption of large consumers, defined as high-rise residential buildings. Water consumption data from the past three years in São Paulo was obtained with the support of Sabesp, and 21 building characteristics are integrated. The dataset includes 3,299 buildings, covering 276,670 apartments. The results may support WD simulations, with potential Smart City applications such as leakage detection and monitoring population density shifts and their impacts on water consumption.

The research hypothesis is that building characteristics provide strong descriptive and predictive power, reducing reliance on population, demographic, and income estimates, which may carry uncertainties and potential biases. After data integration and preprocessing, the most relevant variables for the water consumption in São Paulo are explored. An estimation of the consumption in high-rise buildings is presented, along with an analysis of the prediction errors.

2 Related Work

2.1 Related Studies on Water Comsumption

The related studies found by the authors investigate the demand profile of whole cities [10,13,14], explore and define consumption profiles, based on either hourly/daily patterns [13,15,16,20,22,24] or appliance usage [17,19,27,34]. Other works explored the average consumption of dwellings [4,11,21,31] or per building [5,9,29], with the latter being the focus of this study.

Regarding the techniques used, identification of statistical correlation, distribution analysis, and significance tests are employed to identify relevant variables. To estimate consumption, multiple linear regressions are commonly used [6,11,13,24,29,31,33]. The use of this easily explainable models is particularly interesting, as they provide coefficients that help assess the contribution of each variable to water consumption. GIS techniques [3,4,13,15,16,21,23,24,30] were also widely used, particularly considering Moran's I statistic and Spatial Lag/Correlation, in order to analyze spatial dependencies.

2.2 Variables used to Explaining Urban Water Consumption

The selected variables are grouped into different domains. A subset of the variables proposed in the literature was used, since some of them are difficult to collect and may be unfeasible to collect, such as considering the type of plants in gardens [6]. Other variables may be biased, such as differentiating foreigners from residents [30,33] or considering religion [1].

Climatic Variables: In the context of climate data usage, it is common to find studies that investigate relationships with temperature, humidity, wind, and precipitation [10,14,22,31,12,16], as well as indirect influences through the presence of rainwater reuse systems [34] or alternative water supply systems [8]. Correlations with thermal sensation have also been observed [10,31].

Analyzing data from 38 Chinese cities [16], it was concluded that availability, tariff pricing, and the adoption of water-saving technologies become more significant compared to climatic factors. Notably, when consumers were grouped and a regression was performed [22], building characteristics and consumption related variablessuch as hourly consumption, previous week's consumption at the same hour, and day of the weekwere more influential than temperature.

Variables Related to Appliances: There are uncertainties regarding the consideration of equipment consumption (e.g., faucets, showers). Different usage habits [6,8,11,17,19,20,27,34] have been observed, aiming to define hourly consumption profiles. Some influencing factors include the age of residents [6], habits such as cooking at home versus dining out [20], the presence of efficient appliances [34], or the use of water reuse systems [8,17].

A challenge in considering appliance consumption is the need for *in loco* measurements [19,34] by residents, or the use of estimatives. Additionally, cultural specificities exist, such as in India (the most populous country), where 67% of the population uses traditional bucket baths instead of showers [25].

Population Variables: Common considerations include population estimates [10,19,20,12,23,31,33], the urban/rural resident ratio [13,16,31], and population density [13,24,30,31,33]. The distinction between daytime and nighttime inhabitants [26] and the proportion of people per household, room, or bathroom [24,29] are also used. A potential limitation is the necessity of conducting surveys and questionnaires, which restricts both reach and the number of observations.

A widely recognized finding is that the higher the number of residents in a household, the lower the *per capita* water consumption [6,8,11,26,27,31,33,34]. This is expected, as the property itself remains the same, and when more people share it, the "maintenance cost" is distributed among them [26,27,31].

Census data is commonly used, although it is sometimes outdated [3,33]. A notable observation, as pointed out by [32] for Spain's coastal region, is that cities with high tourism influx, seasonal population variations, or large number of short-term rental properties, can distort data. Only [15,32] consider this factor, which can present challenges and limitations in certain locations.

Consumption as a Variable: The variability of consumption depending on temporal factors, such as the day of the week and time of day, has been observed [14,15,20,22], as well as a reduction in consumption during nighttime hours [14,15,20]. Regarding the inclusion of property-related variables, [22] argues that they are not highly effective, but it is important to note that their approach relies on smart metering for determining hourly consumption.

In Brazil, [5] identifies the regression components that most contribute to estimating consumption as sewage collection and alternative water sources (for estimating property-level consumption), with piped water access also being considered [4,23]. The effects of water supply interruptions have been recognized as an important variable [4,8,24] in studies conducted in Mexico City, Mexico.

Demographic Variables: These are widely considered [6,8,15,25,26,27]. Although they represent universal characteristics, gender, age, type of employment among adults, and level of formal education exhibit different distributions de-

pending on the level of development of a city or country. More developed regions face an aging population and generally provide higher levels of formal education.

Additionally, there is evidence of potential negative bias propagation, such as the questionable use of the female gender to explain increased consumption. There is assumptions that women spend more time at home [6], caring for the household and children [11]. Arguing that areas with a higher male-to-female ratio tend to have higher income levels, [31] uses gender as a proxy for income.

Financial Variables: The integration with the financial domain is widely explored, with the most common approach being the use of residents' income [4,6,8,11,19,20,25,26,27,31,34]. The value of the tariff charged for consumption is also considered [6,15,16,31], as well as property ownership [27,8], Gross Domestic Product (GDP) [16,12,24,31], and property prices [22,25,29].

In Brazil, [3] conducts a spatial analysis in Campina Grande, finding differences of more than 10% between low- and high-income neighborhoods. A similar approach is used by [21], on a national level, focusing on the city of Fortaleza. Both studies conclude that income inequality is reflected in water consumption, with wealthier areas consuming more.

In Aveiro, Portugal, [27] observes a statistical difference between the lowest income group and others, particularly in households with 3-4 people, where consumption in the lowest income group is approximately 37% lower. Finally, in Seville, Spain, [33] uses the property tax as a proxy for income. It is observed that high-income populations revitalize areas previously occupied by low-income groups, specifically in central regions, while the poor move further away. This situation affects the dynamics of the WD within the city.

Building Variables: Various studies aim to relate the characteristics of buildings to explain water consumption by their residents. The most commonly explored conditions include the number of rooms, especially bedrooms and bathrooms [6,8,11,25,29], presence of a swimming pool [6,8,29], constructed area [11,22,29,30,33], age of the building [8,22,27,29], and building type (single-story or vertical) [8,26,29,33].

According to [8,11,30], built area is typically associated with higher consumption levels, and [26] finds a difference of over 30% between properties larger than $100m^2$ and those smaller than $50m^2$. Also, the number of bedrooms and bathrooms figures as two of the most important variables [6,8,11,25,29]. The first reflects an estimate of how many people occupy the property, and the second is one of the areas linked to water consumption.

Regarding building characteristics, [27] finds no impact on consumption when considering the age of the building in Aveiro, Portugal. However, in Joinville, Brazil, [8] concludes that older buildings tend to have higher consumption, a hypothesis related to damaged pipes, lack of sustainable technologies.

In summary, the variables considered and the focus of the research are compiled in Table 1. The themes of the research include Per Capita Consumption (Cpc), Household Water Consumption (HWC), Water Demand Profiling of Cities (WDP-C), Consumption Profiling of Households (CP-H), Consumption

Profiling per Capita (CP-pc), Water Demand Distributions (WD-D), and Water End Use (WEU).

Table 1: Taxonomy of Studies on Water Consumption and Socio-Economic Factors. The highlighted variables are the most important in the study. Exploratory analysis are highlighted in purple, while regression tasks are highlighted in green. **Smp** Sample size; Domains: **Clm** Clima, **Dem** Demographic, **Bld** Building, **Ppl** Populational, **Eqp** Equipment, **Fnn** Financial, **Cns** Consumption

Ref	**Focus**	**Smp**	**Source**	**Clm**	**Dem**	**Bld**	**Ppl**	**Eqp**	**Fnn**	**Cns**
[25]	WEU	248	Survey		✓	✓			✓	✓
[30]	WD-D	-	BD	✓		✓	✓			
[14]	WDP-C	-	Meter	✓						✓
[10]	WDP-C	-	Meter	✓			✓			
[12]	WDP-C	-	BD	✓			✓		✓	
[23]	WDP-C	-	BD				✓			✓
[15]	WDP-C	-	BD		✓				✓	✓
[22]	CP-H	90	Meter	✓		✓				✓
[31]	Cpc	-	BD	✓	✓	✓	✓	✓	✓	✓
[6]	Cpc	532	Survey		✓	✓		✓	✓	
[34]	Cpc	151	Survey			✓		✓	✓	
[26]	Cpc	900	Survey		✓	✓		✓	✓	
[8]	**HWC**	108	Survey		✓	✓		✓	✓	
[19]	WEU	36	Survey				✓	✓	✓	
[17]	WEU	48	Survey					✓	✓	
[24]	WD-D	-	BD			✓			✓	✓
[3]	WD-D	-	BD	✓	✓		✓	✓	✓	
[20]	CP-pc	36	Survey					✓		
[33]	Cpc	-	BD		✓	✓	✓			
[27]	Cpc	53	Survey		✓	✓			✓	✓
[21]	Cpc	-	BD						✓	
[16]	Cpc	-	BD	✓			✓	✓	✓	✓
[4]	Cpc	-	BD				✓		✓	✓
[11]	HWC*	380	Survey		✓	✓		✓	✓	
[9]	**HWC**	394	Survey		✓	✓	✓	✓	✓	
[29]	**HWC**	78	Survey			✓	✓			✓
[5]	**HWC**	89	Survey			✓	✓			✓
Ours	**HWC**	3299	BD			✓			✓*	

In the present analysis, data characterizing vertical residential buildings are obtained. The reason for the asterisk in the Financial domain in Table 1 is due to the consideration of the price and construction standards of the apartments and land, which can serve as a proxy variable for the residents' purchasing power.

There is a predominance of the financial domain (18) and building construction characteristics (15), indicating good coverage in the literature on related

topics. Despite this, few recent studies focus on estimating urban domestic water consumption. Some distinguishing aspects of the present study include the number of collected observations and its specific focus on high-rise buildings.

3 Water consumption and buildings characteristics

3.1 Data Acquisition and Integration

With the support of Sabesp, consumption data for the last three years in the city of São Paulo is obtained. Integration is performed with public municipal property tax data (IPTU) and data from Embraesp (Empresa Brasileira de Estudos de Patrimônio). The Embraesp database contains building characteristics such as the number of blocks, elevators, penthouses, apartment types, and floor levels. It also includes unit attributes such as the number of bedrooms, bathrooms, parking spaces, square footage, and price (adjusted to December 2024 values). From the IPTU database, data on land size and price, built area, and price per square meter, as well as the building quality standard, are obtained.

The lack of standardization between adresses is a challenge. The Sabesp data includes the provisional address of the building, the Embraesp data is the result of data collection, and the IPTU database contains official data. Furthermore, there are cases (Figure 1) that more than one numbering was adopted, for convenience or practicality. Given the name of a street or ZIP code, variations in numbering are accepted within a margin (related to the land frontage).

Fig. 1: Problematic case in database integration - Google Earth/Maps

To validate the data, the building's age, number of floors, number of apartments, and land size with street frontage are cross-referenced. Finally, a web-scraping approach using Google Maps is implemented to resolve discrepancies in cases with multiple possibilities, using the building's name (when available).

3.2 Challenge in Consumption Estimation

It is assumed that large consumers, such as vertical residential buildings, have a non-negligible impact on the urban water distribution. In this scenario, data from the municipal tax database is used to assess the trend of the built area of new apartments in São Paulo, Brazil. Consumption estimation (m^3/month/building) is performed using Equation 1 from Sabesp's technical standard NTS181 [28]. The Mean Absolute Error (MAE) between the estimative and the real consumption is calculated, as shown by the dashed line in Figure 2.

$$Consumption = -21.1 + 0.0177 \cdot A + 2.65 \cdot B + 3.97 \cdot D - 50.2 \cdot P_D + 46 \cdot V_G \quad (1)$$

The monthly *Consumption* (m^3/month/building) is related to the Constructed Area A, number of Bathrooms B, and Bedrooms D, where P_D is 1 if $D > 3$, and 0 otherwise, as well as the number of Garage Spaces per apartment V_G.

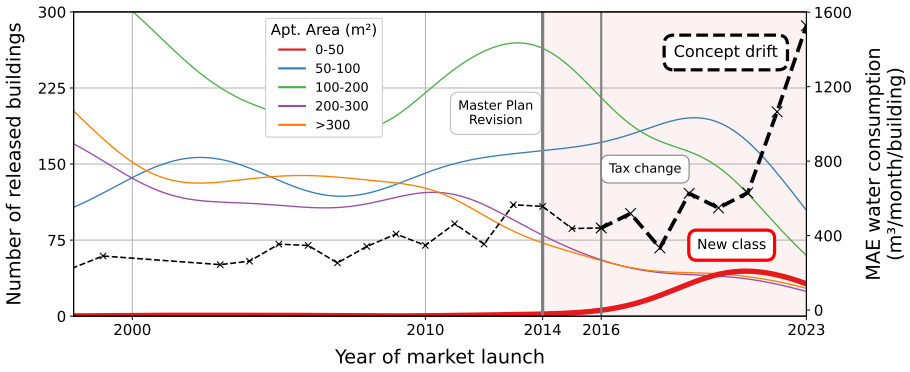

Fig. 2: Relationship between Real Estate Developments and Concept Drift in Sabesp's Regressor Building Data (IPTU) and Consumption Data (Sabesp)

In 2014, the city master plan started encouraging the construction of small-sized apartaments near public transport. In 2016, the local government issued a decree that altered the classification of these properties, which in turn affected the taxes paid. This legislative change gave rise to a new class of properties with up to $50m^2$ of constructed area, that affects the consumption estimation.

Analyzing the consumption estimation error from Sabesp (dashed line), it is evident that, alongside the emergence of the new class, the average annual errors increase dramatically. This analysis allows for the identification of the effect (explosion in error) and the cause (new class) of a Concept Drift. This scenario motivates the search for a regression model that better explains water consumption using only building-related variables, as in the Sabesp model. The results of this study will support the revision of the NTS181 technical standard.

4 Experiments

For the experiments, a preliminary treatment of the consumption data was carried out. Based on experimental results, thresholds were adopted for the data preprocessing. Outliers corresponding to exceptionally high deviations (18 times) from a clients moving average were removed. To avoid strong variations, a smoothing technique was applied to consumers with a minimum consumption of $2m^3$/month. A smoothing operator was applied using median values over the last five months when consumption exceeds five times the limit.

Regarding outliers in the building features, automating the process and removing them using standard deviation seemed to lose valuable information. Therefore, manual removal was performed on the long tails distributions. For instance, entries with more than $400m^2$ (<1% of cases), values above U$1,000,000.00 (<2%) or U$4,000.00/m^2 (<2%) are removed. These are exceptions in the dataset, and the estimation of high-end properties may be a subject for future studies.

4.1 Exploratory Analysis

To illustrate the relationship between variables and consumption by unit, an experiment using the UMAP is conducted. This technique transforms the data from a high-dimensional space into a lower-dimensional (latent) space. Its key feature is the preservation of neighborhood relationships, meaning points that were close in the original space tend to be neighbors in the latent space.

The attributes selected include the number of bathrooms/bedrooms/garage spaces, the value per square meter of the land, the number of apartments per floor, and the area of the apartments. For comparison, Principal Component Analysis (PCA) is also used, and the results are shown in Figure 3. A good separation between different consumption levels is observed, and taking advantage of this observation, regression techniques are trained in the latent space.

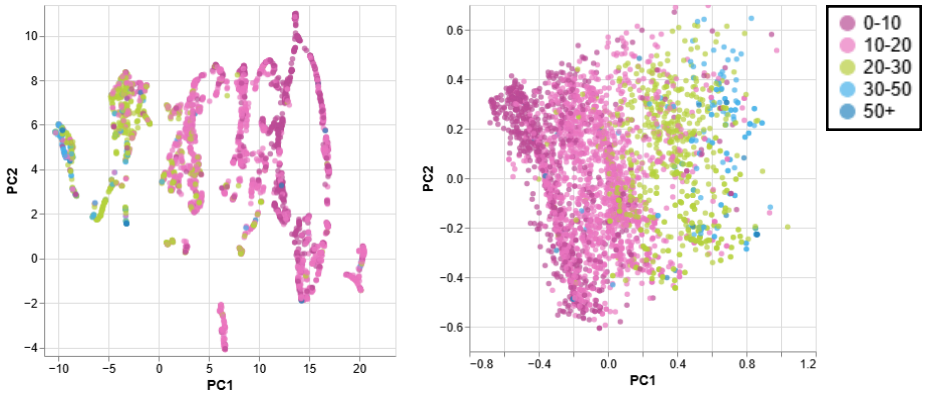

Fig. 3: Latent space - UMAP Latent space - PCA

To analyze the relationship between multiple variables and consumption, parallel coordinate visualization is used. Min-Max normalized values are employed, meaning values close to 1 represent the maximum and values close to 0 represent the minimum for each variable. As shown in Figure 4, for the top 25% highest consumers, there is an almost direct relationship between consumption and the number of bedrooms, bathrooms, and garage spaces, as well as the area and price of the apartment. It is worth noting that the land area and units per floor are associated with the bottom 20%. In other words, the larger consumers are buildings with relative few units, large square footage, and small plots of land.

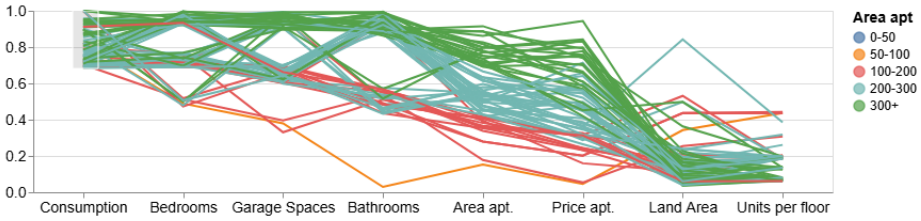

Fig. 4: Relationship between consumption and 7 selected variables

In the distributions shown in Figure 5, the situation is similar to what has been observed in the literature, with a positive correlation between water consumption and the size of the residence [8,11,26,30] as well as the number of bathrooms [6,8,11,25,29]. This characterization is important because, as already observed, changes are occurring in the real estate market of São Paulo. It is noteworthy that apartments up to 50m^2 have significantly lower consumption.

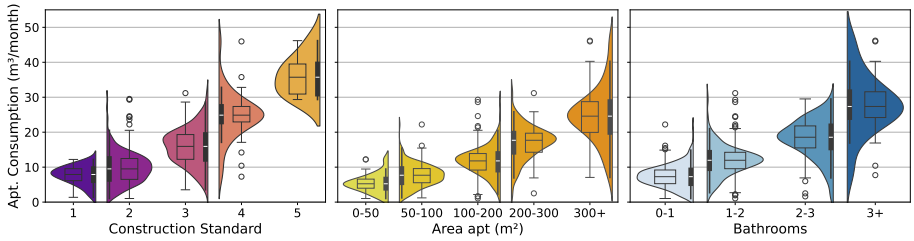

Fig. 5: Consumption distributions related to different variables

Furthermore, the financial variable related to the construction standard (CS) shows that units with CS 1 and 2 (more modest) have a consumption of around 10m^3/month, while units with high standards consume nearly three times as much. A positive correlation between consumption and income has been observed [3,8,21], and the CS may serve as a proxy for the residents' purchasing power.

4.2 Apartment Consumption Estimation

The dataset is splited into training and testing sets in an 80-20 ratio, and the data is normalized using the MinMax technique. The only categorical variable is the Construction Standard of the property, with 5 distinct classes, and Ordinal Encoding is applied. All implementations used are from the Sklearn library.

Different variable sets are tested, ranging from 4 to 21 variables. For feature selection, Recursive Feature Elimination (RFE) is used. In this method, the importance of each variable in the set is calculated, and the least important variables are recursively eliminated until the desired number is reached.

Fourteen different algorithms are tested, including Linear Regression and the regularized variants Ridge and Lasso Regression. Also tested are cases that transform the data into a new latent space and then apply regression, such as Principal and Independent Component Regression. The first uses Principal Component Analysis (PCA) to create the latent space, while the second uses Independent Component Analysis (ICA). One difference of ICA is the relaxation of the orthogonality condition of PCA when creating new components.

Various tree-based strategies are chosen, including Decision Tree (DT) and ensemble tree methods such as Random Forest and Extremely Randomized Trees. The use of DT with boosting strategies is employed with Gradient Boosting and AdaBoost. Finally, algorithms that use distances, such as K-Nearest Neighbors, or hyperplanes, such as Support Vector Machine, are also chosen.

Hyperparameter optimization is performed using grid search. Instead of splitting the training set into training and validation sets, 5-fold cross-validation is used for optimization, with Negative Mean Absolute Error adopted as the scoring parameter (the implementation always aims to maximize the score).

To illustrate the methodology, Figure 6 compares four cases: (i) without data preprocessing, outlier removal, feature selection and regressor optimization (Baseline); with preprocessing the consumption data (anomaly detection and smoothing) and (ii) using only the Sabesp model variables (Sabesp) or (iii) using all available features (All feats); and (iv) with preprocessing and manually removing outliers, feature selection using RFE and hyperparameter optimization using GridSearch (Optimized).

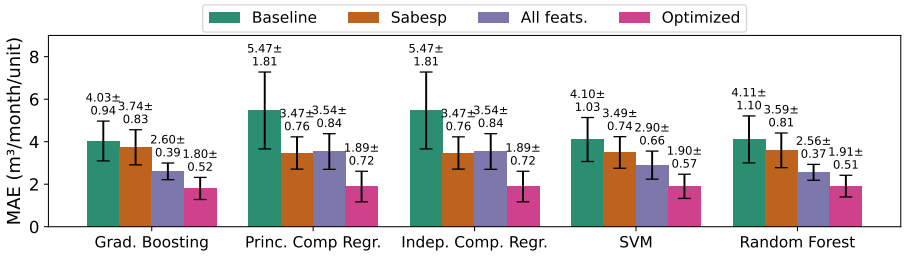

Fig. 6: Comparisson between regressors and methods

Preprocessing yields benefits mainly for linear regressors (PCR and ICR). Notably, Sabesp variables have good prediction power, but there is improvement when all features are included. Manual outlier removal, combined with feature selection and hyperparameter optimization, leads to considerable gains across all cases. Sensitivity analyses on outlier removal will be explored in future work.

The evaluation of the results is carried out using 10-fold cross-validation. The metrics considered, aiming for comparability with previously cited studies, include Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Rš. The regression algorithms used are listed in Table 2, along with the evaluation metrics.

Table 2: Regression Results. Highlighted in green the Tree models and in purple the Linear Regression models - limited to the top 5

Model	MAE	MAPE	RMSE	Rš
Gradient Boosting **(GB)**	$\mathbf{1.80 \pm 0.52}$	0.22 ± 0.08	$\mathbf{2.31 \pm 0.91}$	$\mathbf{0.68 \pm 0.25}$
Principal Component **(PC)**	1.89 ± 0.72	$\mathbf{0.21 \pm 0.09}$	2.54 ± 1.02	0.61 ± 0.33
Independent Component **(IC)**	1.89 ± 0.72	$\mathbf{0.21 \pm 0.09}$	2.54 ± 1.02	0.61 ± 0.33
Support Vector Machine	1.90 ± 0.57	0.23 ± 0.08	2.59 ± 0.82	0.62 ± 0.27
Random Forest	1.91 ± 0.51	0.25 ± 0.10	2.60 ± 0.86	0.59 ± 0.36
Huber Regression	1.92 ± 0.80	0.22 ± 0.10	2.54 ± 1.06	0.61 ± 0.33
Extr. Trees	1.93 ± 0.66	0.24 ± 0.11	2.67 ± 1.12	0.59 ± 0.33
Ridge Regression	1.94 ± 0.73	0.23 ± 0.09	2.58 ± 1.04	0.61 ± 0.32
Bayesian Ridge Reg.	1.95 ± 0.81	0.22 ± 0.10	2.60 ± 1.10	0.60 ± 0.36
AdaBoost	1.96 ± 0.45	0.25 ± 0.11	2.63 ± 0.79	0.59 ± 0.33
K-Nearest Neighbors	2.02 ± 0.49	0.25 ± 0.09	2.83 ± 0.85	0.57 ± 0.23
Linear Regression	2.03 ± 0.86	0.23 ± 0.11	2.71 ± 1.09	0.57 ± 0.35
Lasso Regression	2.09 ± 0.57	0.25 ± 0.07	2.83 ± 0.91	0.57 ± 0.28
Decision Tree	2.17 ± 0.76	0.27 ± 0.09	2.80 ± 0.89	0.51 ± 0.38

The use of trees and boosting showed better performance, as well as the strategy of creating a latent space and then performing linear regression. Significance testing between GB and PC/IC is conducted pairwise using the Wilcoxon Signed-Rank Test. A statistical difference (p-value < 0.01) is found for MAE, RMSE, and Rš, indicating that GB outperformed the others.

A residual analysis is performed. Normality tests are conducted using the Kolmogorov-Smirnov and Shapiro-Wilk tests, and homoscedasticity is assessed using the Breusch-Pagan Test. At a 5% significance level, normality and homoscedasticity are confirmed. By selecting the largest errors (Figure 7), it is observed that they are related to apartments with few bathrooms and small built areas, but with high density (total number of units in the building).

It is worth noting the transformation applied to the target variable. To estimate total building consumption and compare it with the Sabesp regressor MAE, better performance was achieved by modeling consumption per unit rather than

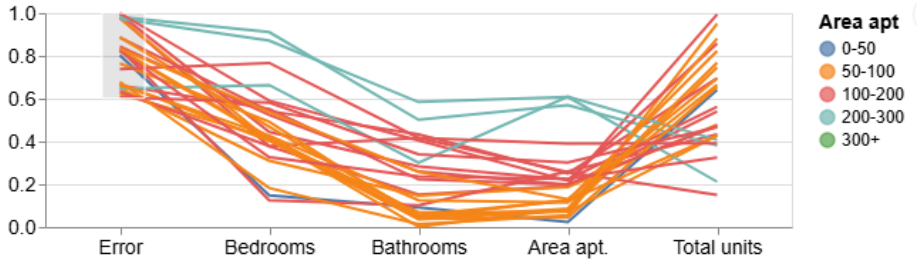

Fig. 7: Relationship between regression residuals and selected variables

for the entire building. In older buildings without individual metering, only total consumption is available, and average per-unit consumption is estimated. Where individual measurements exist, the mean of actual unit consumption is used for estimation. For comparison with Sabesp's regressor, individual units errors are summed, and the Mean Absolute Error is computed at the building level.

5 Discussion

The careful cross-referencing of datasets allows the data from Sabesp to be linked with property characteristics with a high degree of reliability, given the checks described. In this context, integrating data from IPTU and Embraesp proves extremely fruitful, providing important variables related with consumption.

The relative ease of obtaining reliable data is considered an advantage for this approach and is thus adopted in this research. This statement can be justified by the ease of accessing real estate registry databases, such as those used in the present study, or even collecting data through questionnairesmeasuring the number of rooms/area is considered simpler than measuring the flow rate of individual appliances (faucets, showers, etc.) for each use.

It is noted that selected variables create latent spaces in which the Consumption per unit (in m^3/month) have representations of the classes in well-defined regions for each class. It is observed that variables such as the number of bedrooms and bathrooms, as well as the unit area and land size, are good descriptors of consumption. Naturally, these are attributes correlated with the number of occupants in the property and the number of sanitary appliances, considering that maintenance and cleaning are proportional to the unit and land area.

Across the top five modelsGB, PCR, ICR, SVM, and RFseven features were consistently selected: number of Bedrooms, Bathrooms, and Garage Spaces; Unit Area; Units per Floor; Building Age; and Land Price per m^2. These variables appear to be the most informative building characteristics for predicting water consumption. Notably, the first four are also included in Sabesps regressor.

Estimating monthly consumption in Brazil, [9] uses Univariate Regression Trees with acceptable results, achieving an RMSE of 5.90 m^3/month/household, whereas the present study reached values of 2.31 m^3/month/household. Using

multiple linear regressions, [5] reports a MAPE of 16.78%, considering population variables and the availability of water and sewage services. This may explain why their results were better than the 22% obtained. However, it should be noted the heterogeneity among 3,299 buildings considered in this study.

Also using linear regression, [29] achieves a MAE of 1.23 m^3/month, a value comparable to the 1.80 m^3/month obtained in the present study. Regarding Sabesp's regressor, for the entire building, it presents an average MAE of 620 m^3/month (considering concept drift). Even before the emergence of the new class, the MAE values were around 350-400 m^3/month, and the strategy of estimating the consumption per apartment and then estimating the total consumption resulted in an MAE of 202.46 m^3/month.

6 Conclusion

Population growth puts pressure on the (already degraded) environment. Furthermore, another aspect to consider is the transition of rural populations to urban centers, adding stress to the infrastructure of cities. In this scenario, with the increase in population and the expected concentration of people in urban centers over the coming decades, it is imperative to seek solutions that promote quality of life while preserving natural resources.

Focusing on large water consumers, represented by vertical residential buildings, three datasets are integrated, with real consumption data provided by Sabespthe largest sanitation company in Latin America. Data from 3,299 buildings, encompassing approximately 276,670 apartments, with 21 variables describing the properties, are gathered from the city of São Paulo. A possible concept drift is detected in Sabesp's model, potentially related to changes in the city's master plan and subsequent impacts on real estate development.

The research hypothesis is that only variables describing the characteristics of the buildings are sufficient to generate a good estimate of urban residential consumption in vertical buildings. This is partially verified, with performance similar to models that use population-related variables (population estimates, residents per household, etc.), financial variables (income estimates, GDP, etc.), or demographic variables (age, gender, educational level, etc.).

To confirm this, it is necessary to quantify the impact of considering the variables described above. Additionally, the influence of neighbors can be explored using GIS, and the impact of seasonality and the percentage of vacant units can be examined. These are relevant considerations that could improve models for estimating monthly water consumption, which are topics of ongoing work.

Acknowledgments. The study has the support of the São Paulo Research Foundation (FAPESP), by an undergraduate scholarship for scientific research (Grant 2024/08236-8) and by the Brazilian Applied Research Center on Smart and Sustainable Cities IARA (Grant 2020/09835-1). This study also had the support of the Water and Sewage Company (Sabesp), by providing the data used in the experiments.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahmed Kalifa, Alanoud Al-Maadid, I.K., Makropoulos, C.: Individual water consumption behavior in relation to urban residential dynamics: The case of qatar. Urban Water Journal **18**(10), 806–816 (2021). https://doi.org/10.1080/1573062X.2021.1941135
2. Alves, M., Appert, M., Montès, C., Tapia Martín, C.: Producing and living the high-rise New contexts, old questions? Series in Built Environment Bridging Languages and Scholarship, pp. 175–196 (03 2024)
3. Brito, H.C.d., Rufino, I.A.A., Barros Filho, M.N.M., Meneses, R.A.: Use of spatial data in the simulation of domestic water demand in a semiarid city: The case of campina grande, brazil. Urban Science **7**(4) (2023). https://doi.org/10.3390/urbansci7040120
4. Carolina Massiel Medina-Rivas, Jorge Armando Morales-Novelo, L.R.T., Revollo-Fernández, D.A.: Mexico citys decline in per capita domestic water use: a comprehensive spatial-temporal study. Urban Water Journal **22**(1), 1–15 (2025). https://doi.org/10.1080/1573062X.2024.2423400
5. Dias, T.F., Kalbusch, A., Henning, E.: Factors influencing water consumption in buildings in southern brazil. Journal of Cleaner Production **184**, 160–167 (2018). https://doi.org/10.1016/j.jclepro.2018.02.093
6. Domene, E., Saurí, D.: Urbanisation and water consumption: Influencing factors in the metropolitan region of barcelona. Urban Studies **43**, 1605–1623 (08 2006). https://doi.org/10.1080/00420980600749969
7. DOttaviano, C., Bossuyt, D.M.: Vertical incremental housing in são paulo. the case of minha casa minha vida entidades. International Journal of Housing Policy **0**(0), 1–26 (2024). https://doi.org/10.1080/19491247.2024.2308716
8. Garcia, J., Salfer, L.R., Kalbusch, A., Henning, E.: Identifying the drivers of water consumption in single-family households in joinville, southern brazil. Water **11**(10) (2019). https://doi.org/10.3390/w11101990
9. Grespan, A., Garcia, J., Brikalski, M.P., Henning, E., Kalbusch, A.: Assessment of water consumption in households using statistical analysis and regression trees. Sustainable Cities and Society **87**, 104186 (2022). https://doi.org/10.1016/j.scs.2022.104186
10. Haque, M.M., de Souza, A., Rahman, A.: Water demand modelling using independent component regression technique. Water Resources Management **31**(1), 299–312 (Jan 2017). https://doi.org/10.1007/s11269-016-1525-1
11. Iman Alharsha, Fayyaz A Memon, R.F., Hussien, W.A.: An investigation of domestic water consumption in sirte, libya. Urban Water Journal **19**(9), 922–944 (2022). https://doi.org/10.1080/1573062X.2022.2105239
12. Joseph, N., Ryu, D., Malano, H.M., George, B., Sudheer, K.P.: Estimation of statewide and monthly domestic water use in india from 1975 to 2015. Urban Water Journal **18**(6), 421–432 (2021). https://doi.org/10.1080/1573062X.2021.1893362
13. Kassay, A.B., Tuhar, A.W., Ulsido, M.D.: Integrated modelling techniques to implication of demographic change and urban expansion dynamics on water demand management of developing city in lake hawassa watershed, ethiopia. Environmental Research Communications **5**(5), 055012 (may 2023). https://doi.org/10.1088/2515-7620/acd512
14. Kavya, M., Mathew, A., Shekar, P.R., P, S.: Short term water demand forecast modelling using artificial intelligence for smart water management. Sustainable

Cities and Society **95**, 104610 (2023). https://doi.org/10.1016/j.scs.2023.104610
15. Loureiro, D., Coelho, S.T., Machado, P., Santos, A., Alegre, H., Covas, D.: Profiling Residential Water Consumption, pp. 1–18. https://doi.org/10.1061/40941(247)44
16. Lu, S., Gao, X., Li, W., Jiang, S., Huang, L.: A study on the spatial and temporal variability of the urban residential water consumption and its influencing factors in the major cities of china. Habitat International **78**, 29–40 (2018). https://doi.org/10.1016/j.habitatint.2018.05.002
17. Marinoski, A.K., Vieira, A.S., Silva, A.S., Ghisi, E.: Water end-uses in low-income houses in southern brazil. Water **6**(7), 1985–1999 (2014). https://doi.org/10.3390/w6071985
18. Marques, E., Minarelli, G.: Verticalization and residential affordability in são paulo. Available at SSRN 4930980
19. 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 **466-467**, 467–474 (2014). https://doi.org/10.1016/j.scitotenv.2013.07.036
20. Matos, C., Teixeira, C.A., Duarte, A., Bentes, I.: Domestic water uses: Characterization of daily cycles in the north region of portugal. Science of The Total Environment **458-460**, 444–450 (2013). https://doi.org/10.1016/j.scitotenv.2013.04.018
21. Tereza Margarida Xavier de Melo Lopes, Samiria Maria Oliveira da Silva, L.d.S.S., Soares, R.B.: Water and socioeconomic inequalities: spatial analysis of water consumption in brazil. Urban Water Journal **21**(9), 1056–1070 (2024). https://doi.org/10.1080/1573062X.2024.2397791
22. Pesantez, J.E., Berglund, E.Z., Kaza, N.: Smart meters data for modeling and forecasting water demand at the user-level. Environmental Modelling & Software **125**, 104633 (2020). https://doi.org/10.1016/j.envsoft.2020.104633
23. Ramos-Bueno, A., Galeana-Pizaña, J.M., Perevochtchikova, M.: Urban water consumption analysis through a spatial panel modeling approach: a case study of mexico city, 20042022. Water Supply **24**(9), 3179–3195 (2024). https://doi.org/10.2166/ws.2024.191
24. Ramos-Bueno, A., Perevochtchikova, M., Chang, H.: Socio-spatial analysis of residential water demand in mexico city. Tecnología y ciencias del agua **12**(2), 59110 (2021). https://doi.org/10.24850/j-tyca-2021-02-02
25. Ramsey, E., Berglund, E.Z., Goyal, R.: The impact of demographic factors, beliefs, and social influences on residential water consumption and implications for non-price policies in urban india. Water **9**(11) (2017). https://doi.org/10.3390/w9110844
26. Rondinel-Oviedo, D.R., Sarmiento-Pastor, J.M.: Water: consumption, usage patterns, and residential infrastructure. a comparative analysis of three regions in the lima metropolitan area. Water International **45**(7-8), 824–846 (2020). https://doi.org/10.1080/02508060.2020.1830360
27. S. Costa, I.M., Sousa, V.: Understanding residential water demand: insights from a survey in a mediterranean city. Urban Water Journal **21**(4), 521–537 (2024). https://doi.org/10.1080/1573062X.2024.2312501
28. Sabesp: Norma Técnica Sabesp NTS 181, revisão 4 edn. (novembro 2017)
29. Kairo Pereira Teodoro da Silva, Andreza Kalbusch, E.H., Menezes, G.A.L.: Modeling water consumption in multifamily buildings: a case study in southern brazil. Ur-

ban Water Journal **18**(10), 783–795 (2021). https://doi.org/10.1080/1573062X.2021.1934040
30. Tayebi, S., Feizizadeh, B., Esfandi, S., Aliabbasi, B., Ali Alavi, S., Shamsipour, A.: A neighborhood-based urban water carrying capacity assessment: Analysis of the relationship between spatial-demographic factors and water consumption patterns in tehran, iran. Land **11**(12) (2022). https://doi.org/10.3390/land11122203
31. Fabiano da Veiga, A.K., Henning, E.: Drivers of urban water consumption in brazil: a countrywide, cross-sectional study. Urban Water Journal **20**(10), 1462–1470 (2023). https://doi.org/10.1080/1573062X.2022.2041049
32. Villar-Navascués, R.A., Pérez-Morales, A.: Factors affecting domestic water consumption on the spanish mediterranean coastline. The Professional Geographer **70**(3), 513–525 (2018). https://doi.org/10.1080/00330124.2017.1416302
33. Villarín, M.C.: Methodology based on fine spatial scale and preliminary clustering to improve multivariate linear regression analysis of domestic water consumption. Applied Geography **103**, 22–39 (2019). https://doi.org/10.1016/j.apgeog.2018.12.005
34. Willis, R.M., Stewart, R.A., Giurco, D.P., Talebpour, M.R., Mousavinejad, A.: End use water consumption in households: impact of socio-demographic factors and efficient devices. Journal of Cleaner Production **60**, 107–115 (2013). https://doi.org/10.1016/j.jclepro.2011.08.006, special Volume: Water, Women, Waste, Wisdom and Wealth