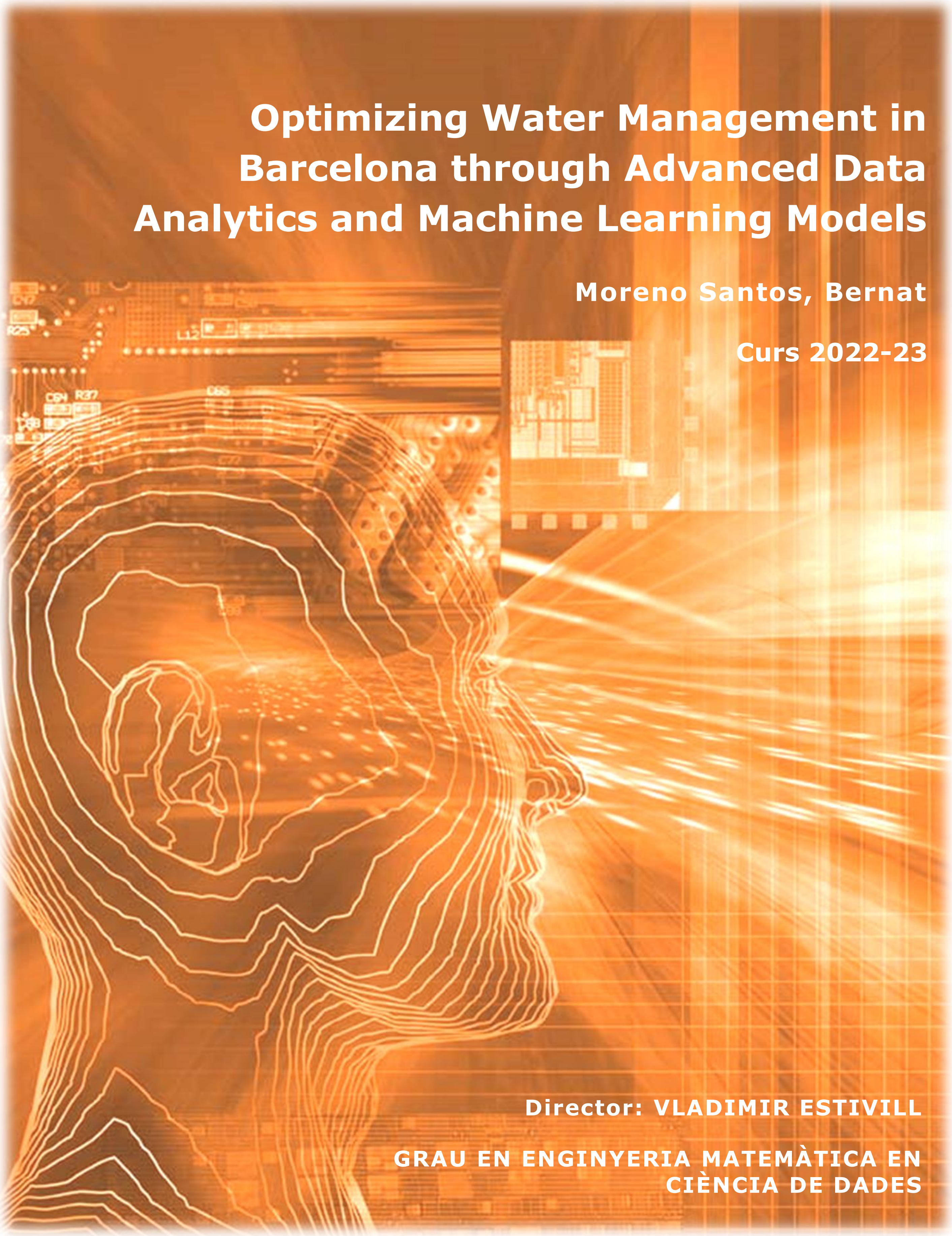


Optimizing Water Management in Barcelona through Advanced Data Analytics and Machine Learning Models

Moreno Santos, Bernat

Curs 2022-23

Director: VLADIMIR ESTIVILL

GRAU EN ENGINYERIA MATEMÀTICA EN CIÈNCIA DE DADES

upf. Universitat Pompeu Fabra Barcelona

Escola d'Enginyeria

Treball de Fi de Grau

Dedication

To the past version of myself, who is proud of my current me. To my dear family and in memory of my cherished grandpa.

Acknowledgments

First and foremost, I am deeply pleased with my supervisor, Vladimir Estivill, for his exceptional guidance, expertise and patience.

I am immensely grateful to my friends and classmates who have been a constant source of motivation and inspiration.

I would like to extend my gratitude to the faculty members and researchers who have provided valuable feedback during the course of my study.

Summary

The project involved analyzing domestic water consumption in Barcelona and using machine learning models to forecast time series. It is an extension of the AB Data Challenge, a collaborative initiative between Aigües de Barcelona, universities, and research centers in Catalonia. The goal was to use data collected by Aigües de Barcelona for interdisciplinary innovation. The study utilized geo-temporal data from municipalities in the metropolitan area for 2019 to 2021. Machine learning algorithms, such as k-Means clustering and linear regression, were employed for data analysis and visualization. Evaluation metrics were used to assess the fit of the models. The project provided valuable insights into consumption patterns and the impact of COVID-19 on domestic water consumption. Additionally, explainable artificial intelligence (AI) was used to evaluate demographic, socio-economic and meteorologic factors influencing consumption. The study aimed to forecast water consumption figures using the AutoRegressive Integrated Moving Average (ARIMA) Model and compared the performance of several machine learning models. The focus of the project was to utilize this data to drive innovation and develop solutions to meet the changing needs of the Barcelona metropolitan area.

Resum

Aquest projecte ha consistit a realitzar un anàlisi de dades de consum domèstic d'aigua a Barcelona per identificar patrons de consum i utilitzar algoritmes d'aprenentatge automàtic per predir sèries temporals de dades. Es tracta d'una extensió de la primera edició de l'AB Data Challenge, una iniciativa impulsada per Aigües de Barcelona en col · laboració amb universitats i centres de recerca de Catalunya. El projecte utilitza dades geotemporals de tots els municipis de l'àrea metropolitana per als anys 2019, 2020 i 2021. Per a l'anàlisi i la visualització de les dades, l'estudi aplica algoritmes d'aprenentatge automàtic, incloent k-Means clustering i regressió lineal, i proporciona mètriques d'avaluació del seu rendiment. Els resultats de l'estudi proporcionen informació valuosa sobre els patrons de consum i l'efecte de COVID-19 en el consum domèstic daigua. El projecte també utilitza intel · ligència artificial (IA) explicable per avaluar els factors demogràfics, socioeconòmics i meteorològics que influeixen en el consum d'aigua. L'estudi pretén pronosticar les xifres de consum d'aigua mitjançant el model AutoRegressive Integrated Moving Average (ARIMA) i comparar el rendiment de diversos models d'aprenentatge automàtic. L'objectiu d'aquest projecte és aprofitar aquestes dades per impulsar la innovació i desenvolupar noves solucions que satisfacin les necessitats canviants de l'àrea metropolitana de Barcelona.

Abstract

This research aims to use machine learning algorithms to predict future water use after analysing water data usage in Barcelona from 2019 to 2021. This study will use geo-temporal data to evaluate consumption trends in several areas and time durations. The main goal is to identify and group trends of users thorugh k-Means clustering and use Shapley values to evaluate the relationships between consumption statistics and parameters like household income, weather, and population density. Also, the research will assess how well the ARIMA model can estimate water use trends and provides a comparative analysis with other models such as Long Short-Term Memory (LSTM) model, Prophet Model or Gradient Boosting Regression Trees (GBRT). My study has delivered findings that suggest how Barcelona's water resources might be handled more effectively and can be applied to other cities with the corresponding adaptations.

Table of contents

List of Figures

List of Tables

1 Motivation

According to the Barcelona's Climate Plan from 2018 to 2030[1], natural water supply is expected to reduce slightly, causing greater concerns about future availability given the expected increase in demand. More specifically, by 2050, a 12% decrease in surface resources and a 9% decrease in underground resources are expected, along with a 4% increase in demand for various applications. As a result, the metropolitan area will have a general need for additional potable water resources of 34 hm^3 per year, with Barcelona's need predicted to be 18 hm^3 per year. This rise in demand is expected to accentuate already existing water management issues and lead to future water shortages.

Barcelona and its metropolitan area are home to the majority of the population and economic activities, and their own water resources are not sufficient for their potable water needs. As a result, currently a major portion of the city's water supply derives from the surface resources of other basins. Other sources of supply include subsurface resources, a desalination facility, and rainwater[1].

In view of the current shortage of water resources, during May 2023, the Catalan Water Agency (ACA)[2] decreed that the Ter-Llobregat basin, which feeds the city of Barcelona, was in an exceptional hydrological situation, which means that the city's Drought Action Protocol is now in an exceptional phase. The protocol establishes specific actions to save drinking water in services or activities such as green spaces, ornamental fountains, urban cleaning, sports facilities, swimming pools and vehicle cleaning, with the aim of reducing normal water consumption by 15 % in domestic, commercial, industrial and other recreational use and by 50 % in recreational use involving irrigation.

The 2016 Barcelona Health Survey from Barcelona Public Health Agency (ASPB)[3], points out that there is a big difference in income in Barcelona, depending on the district, and that this has consequences for access to water and energy.

The effects of climate change will also lead to changes in energy consumption patterns, with a forecast for less demand for heating , while the demand for water and cooling could rise[1].

Machine learning and data analysis can help address issues related to water consumption by minimizing the costs while operating within the required set of operational constraints [20]. Identifying hidden patterns of use can be also helpful for improving water management, supporting the city' sustainability goals and promoting conservation efforts. If this project were fully successful, the corresponding findings will be relevant for Aigües de Barcelona and policy makers in several areas of improvement, from building a more sustainable management system to building more accurate loyalty and savings campaigns.

[1] Barcelona's Climate Plan from 2018 to 2030: https://www.barcelona.cat/barcelona-pel-clima/sites/default/files/documents/climate_plan_maig.pdf

[2] Portal de la sequera (ACA): https://sequera.gencat.cat/ca/inici/

[3] La Salut a Barcelona 2016 (ASPB): https://www.aspb.cat/wp-content/uploads/2017/11/Informe_Salut_2016.pdf

2 AB Data Challenge

The AB Data Challenge[4] is a program launched by Aigües de Barcelona in collaboration with universities and research centers in Catalonia to make use of the data collected by the company's tele-reading service and transform it into a source of interdisciplinary innovation. The initiative is designed to promote the inclusion of research, information, and findings that allow the beneficial impact of data usage.

The goal of the AB Data Challenge is to leverage data and technology to address new contexts, needs, challenges, and applications within the Barcelona metropolitan area through open innovation and teamwork. The program involves analysing data and drawing conclusions as a team to enable better social and collective decision-making. The participants will solve problems in innovative, ethical, and socially responsible ways through unique projects, and they will carry out initiatives to address sustainability challenges.

To collect data for the AB Data Challenge, Aigües de Barcelona made available to the participating groups the aggregated geo-temporal data for all municipalities in the metropolitan area for the years 2019, 2020, and 2021. The dataset is divided into three separate folders each representing a different type of water consumption: industrial, commercial, and domestic. For each municipality in the metropolitan area, there is a separate `CSV` file containing the date, type of use, daily consumption per water meter in litres, and the ID of the water counter. All the data from all municipalities sum up a total of 136,079,854 records. In addition, there was a separate `CSV` file containing the average water consumption in litres per person and municipality. The schema of the data from Aigües de Barcelona appears in Appendix A.1. The structure of the `CSV` files can be seen in Appendix A.2.

3 About the data

3.1 Data Collection

I employed a rigorous approach by integrating geo-spatial, demographic, meteorological and socio-economic data to enhance the accuracy of its analysis. My expectation is that by utilising these diverse datasets, I would arrive to a comprehensive understanding of the factors influencing water consumption patterns in Barcelona.

The geo-spatial data for the metropolitan area of Barcelona was obtained from a `geojson` file I downloaded from the Institut Cartogràfic i Geològic de Catalunya[5]. The file contains a geometry component describing the shape and the coordinates in latitude and longitude pairs for each municipality. The properties component contains additional attributes associated with the municipality such as the municipality name and its population.

[4] www.abdatachallenge.cat.

[5] Institut Cartogràfic i Geològic de Catalunya. Retrieved 25/04/2023 from https://analisi.transparenciacatalunya.cat/Urbanisme-infraestructures/Municipis-Catalunya-Geo/9aju-tpwc

Additionally, the water consumption data is supported with demographic information provided by the Institut d'Estadística de Catalunya[6]. Each municipality is backed with data such as number of inhabitants and population density.

Meteorological data attributes are collected from four different weather stations of Xarxa d'Estacions Meteorològiques (XEMA)[7]: Barcelona - El Raval, El Prat de Llobregat, Sant Cugat del Vallès - CAR and Vilanova del Vallès. For each municipality the data of the nearest weather station is assigned. By considering attributes such as temperature, precipitation, relative humidity, and solar radiation, the project gains insights into the weather-related factors driving water consumption patterns.

Socioeconomic information was also incorporated into the analysis. This information included variables such as average net income per household, which was obtained from the Instituto Nacional de Estadística (INE)[8] and provided in a CSV file.

3.2 Feature Engineering

Any data analysis project must include feature engineering since it includes adding new variables or changing existing ones to increase a model's performance for prediction. The new features are divided in the five categories below.

1. Temporal data

In the present study, I have opted to incorporate additional features denoting the classification of each day as either a weekday or a non-weekday. The purpose behind this decision is to enhance the understanding of water consumption patterns. This additional information could be valuable due to separate into different consumption tendencies, primarily due to contrasting schedules and behavioral patterns between weekdays and weekends

2. Demographic data

The integration of demographic data facilitates a comprehensive analysis of the relationship between population characteristics and water consumption. Number of inhabitants and population density provides further granularity by considering the spatial distribution of the population within the study area. This information aids in recognizing localized patterns and disparities in water consumption, which may arise due to differences in urbanization, land use, or infrastructure development. The data was integrated into the analysis by means of a join operation, using the shared identifier of municipality as the common key

[6] Institut d'Estadística de Catalunya. Retrieved 12/05/2023 from https://www.idescat.cat/indicadors/?id=aec&n=15228&t=202000

[7] XEMA. Retrieved 14/05/2023 from https://www.meteo.cat/wpweb/serveis/cataleg-de-serveis/serveis-oberts/dades-obertes/

[8] INE. Retrieved 17/05/2023 from https://www.ine.es/jaxiT3/Tabla.htm?t=30896&L=0

3. Socioeconomic data

Socioeconomic factors have been widely acknowledged as key determinants of consumer behavior, including water usage. By incorporating net average income per municipality household, we aim to explore the potential relationship between socioeconomic status and water consumption. Income levels can significantly impact individuals purchasing power, lifestyle choices, and water usage practices. The data was combined into the study using a join operation, with the municipality name as the common key.

4. Meteorological data

Weather conditions can significantly influence water consumption patterns. Specifically, temperature and relative humidity can affect human comfort levels, leading to changes in water usage for purposes such as cooling and hydration. Precipitation levels can affect the availability of water resources and outdoor water consumption, while solar irradiation can impact the use of water for landscaping and irrigation purposes. For each municipality the data of the nearest weather station of XEMA was assigned for value imputation.

5. Geo-spatial data

Geo-spatial data facilitates the identification of spatial patterns, trends, and relationships, enabling a deeper analysis of the geographically dependent factors that may influence the water consumption. Moreover, by geographically mapping the data and visualizing it using *Tableau*, helps in the identification of potential spatial clusters.

Appendix A.3 shows a table with the variables I decided to use and their corresponding units.

4 Data Cleaning and Preparation

4.1 Outlier detection and elimination

Outliers are data points that are significantly different from other data points in the dataset. Outliers may appear due to measurement errors, data processing errors, or they may represent unusual data [3].

Outlier elimination is a crucial data preparation step used before modeling. Outliers can skew the data's distribution and may have an impact on the model's performance and accuracy. By eliminating them, we can mitigate the effect of extreme values and the remaining data can more accurately reflect the underlying population, leading to models that are more accurate and trustworthy [2] [3].

In this particular case, the dataset presented a high presence of outliers and inconsistent data, such as extremely high consumption values or negative values. The criteria to remove outliers was the following:

1. Greater than two standard deviations

In a normal distribution, approximately 95% of the data points should be within two standard deviations of the mean. To calculate a confidence interval using a 95% confidence level, we can start by finding the appropriate z-score that corresponds to this level of confidence. For a standard normal distribution, which has a mean of 0 and a standard deviation of 1, the z-score that corresponds to a 95% confidence level is approximately 1.96. However, if we are working with a sample from a normal population with a known standard deviation, we need to adjust our calculation based on the sample size and degrees of freedom. Specifically, we use the t-distribution instead of the standard normal distribution. For a two-tailed test with a 95% confidence level and n degrees of freedom, the t-value is approximately 2. Therefore, if we know the population standard deviation, we can calculate a 95% confidence interval for the population mean by adding and subtracting 2 standard deviations from the sample mean [3]. As a result, data points that deviate from the mean by more than two standard deviations may be classified as exceptional or unusual.

2. Negative values

As consumption cannot be negative, negative numbers make no sense in the context of user consumption data. Negative numbers may occur because of miscalculations or other problems with the data processing, and including them in the analysis would result in inaccurate findings. As a result, eliminating negative values guarantees that the data is accurate and valid [5].

Figure 1 provides a colored representation of the detected outliers within the dataset.

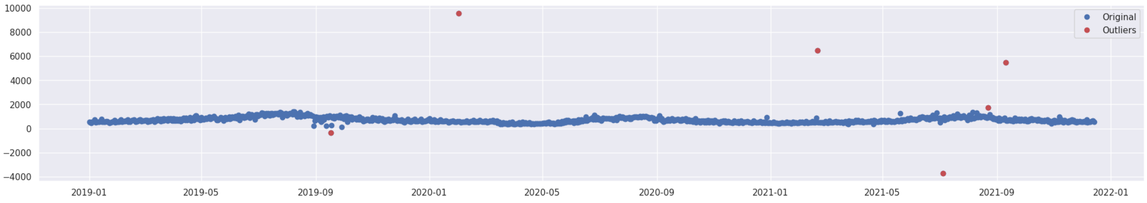


Figure 1: Water consumption with outliers

4.2 Moving Median Smoothing for imputation of missing values

Missing values pose a common challenge in data analysis and can significantly affect the reliability and accuracy of statistical analyses. Various methods have been developed to address the issue of missing data, including imputation techniques that estimate missing values based on available information. Smoothing moving median is a widely used method for imputing missing values in time series with conditional probabilities due to its ability to preserve the overall trend and distribution of the dataset [5].

The smoothing moving median method involves the use of a sliding window to compute the median within the window for each data point. By replacing the original value with the computed median, the method creates a smoothed representation of the dataset. The window size determines the number of adjacent data points considered for calculating the median [7]. The resulting smoothed dataset helps mitigate the impact of outliers and reduces noise, thereby improving the accuracy of subsequent analyses [5].

The smoothing moving median technique can effectively input missing values by using the existing data points in the proximity of the missing value. By replacing the missing value with the median of the neighboring data points, the imputed value aligns with the underlying distribution and trend of the dataset [7]. This imputation approach maintains the original distribution while reducing the bias introduced by alternative imputation methods [4].

The choice of an appropriate window size is crucial in balancing the trade-off between preserving the original data characteristics and smoothing the dataset. Smaller window sizes tend to preserve more details and exhibit a higher sensitivity to outliers, while larger window sizes result in greater smoothing but may obscure important features. Based on prior studies [6] [7], a window size of 4 has been empirically determined as optimal for the smoothing moving median method. This window size strikes a balance between noise reduction and preservation of important trends, providing a reasonable compromise for missing value imputation.

Figure 2 shows a visual representation of the estimated points using the smoothing median technique.

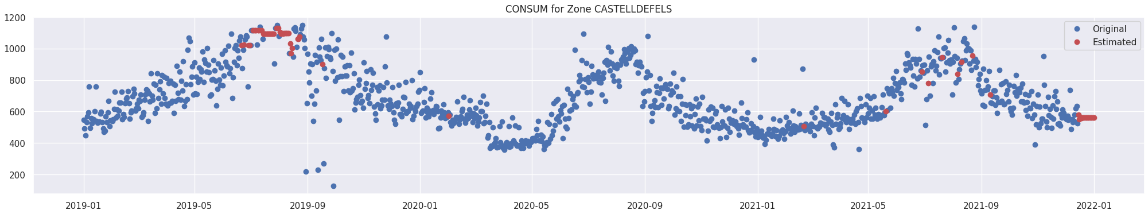


Figure 2: Visual representation of the estimated values

4.2.1 Robustness of median

When dealing with missing values, the choice between using the mean or the median depends on the characteristics of the data. While both methods are common, they are not equally effective in all cases. The mean is suitable for symmetrical data, where values are evenly distributed around the mean [23]. However, when the data is not symmetrical and has outliers, the mean can be significantly influenced by the outliers, leading to a biased estimate of the missing value [1].

The median provides a robust measure of central tendency that is less influenced by extreme values [23]. Huber, P.J. demonstrated that the median is resistant to outliers and provides reliable estimates even in the presence of significant data contamination [24]. Similarly, Rousseeuw, P. J., Hampel, F. R., Ronchetti, E. M., & Stahel, W. A. emphasized the robustness of the median, highlighting its ability to yield accurate results when confronted with outlying observations [25]. These findings establish the superiority of the median as a measure of central tendency in the presence of outliers and support its application in robust statistical analyses.

5 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an approach in data analysis that focuses on gaining an initial understanding and insight into the characteristics and patterns present in a dataset. Due to the unlabeled nature of the data, this chapter is focused on clustering analysis. Clustering is an unsupervised learning approach particularly suited to situations where the nature of the data is unknown or complex, making it an invaluable tool in various fields such as marketing, biology, and social sciences [11]. In contrast to confirmatory data analysis, which focuses on testing pre-existing hypotheses and validating theories, cluster analysis provides a flexible and hypothesis-generating approach that facilitates a comprehensive understanding of the data's underlying structure.

5.1 k - Means Clustering

Clustering is a fundamental data analysis technique used to identify groups or clusters of similar data points based on their intrinsic characteristics [11] [10]. It plays a crucial role in uncovering hidden patterns and structures within datasets, enabling researchers to gain valuable insights and make informed decisions. This section focuses on the application of the k-Means clustering algorithm. By employing k-Means clustering we aim to identify distinct consumption types within a single user's water usage data and group similar consumption patterns within different users.

Let $X = \{x_1, x_2, \ldots, x_n\}$ denote a set of n-dimensional data points, where each point $x_i \in \mathbb{R}^d$ represents a d-dimensional feature vector. The goal of the k-Means algorithm is to partition the data points into k distinct clusters, where k is a user-defined parameter. Each cluster is represented by a centroid μ_j, $j = 1, 2, \ldots, k$, which serves as the representative point for the corresponding cluster [10] [3].

The k-means algorithm aims to find r_{ij} and μ_j that minimize the within-cluster sum of squared distances. The objective function is given as follows [10]:

$$J = \sum_i \sum_j r_{ij} |||x_i - \mu_j|||^2$$

where r_{ij} is an indicator variable defined as:

$$r_{ij} = \begin{cases} 1, & \text{if } x_i \text{ belongs to cluster } j \\ 0, & \text{otherwise.} \end{cases}$$

For the initialization step, we select k initial centroids μ_j, $j = 1, 2, \ldots, k$, either randomly or using a priori knowledge. Then we assign each data point x_i to its nearest centroid using the Euclidean distance metric [10].

On the cluster assignment step, we compute the Euclidean distance from each data point x_i, $\|\|x_i - \mu_j\|\|$, to each centroid μ_j. Assign x_i to the cluster with the nearest centroid by setting $r_{ij} = 1$ for the corresponding j and $r_{ij} = 0$ for $s \neq j$.

For updating the centroid, we recalculate the centroid μ_j of each cluster as the mean of all data points assigned to that cluster: $\mu_j = \frac{1}{|C_j|} \sum_i r_{ij} x_i$, where C_j denotes the set of data points assigned to cluster j.

We repeat cluster assignment step and centroid updating step until convergence is achieved, i.e., when there is no change in cluster assignments [10].

5.2 Comparative Analysis of k-Means and k-Medoids Clustering Methods

k-Means and k-Medoids are both clustering techniques that try to group together data points that have similar characteristics. The method used to determine the cluster centres, or centroids, differs significantly between the two.

The k-Means algorithm aims to minimize the sum of squared distances between data points and their assigned cluster centroids [11]. Recall that the centroid is calculated averaging all the data points in a cluster and that the objective function denoted as $J_{\text{k-means}}$, can be expressed as [10]:

$$J_{\text{k-means}} = \sum_i \sum_j r_{ij} \|\|x_i - \mu_j\|\|^2$$

Unlike k-Means, the k-Medoids algorithm utilizes actual data points, known as medoids, as representatives of clusters. The medoids are the cluster's most centrally situated data point [10] [3]. The objective function for k-medoids aims to minimize the distance $d(x_i, m_j)$ from point x_i to medoid m_j and is denoted as $J_{\text{k-medoids}}$

$$J_{\text{k-medoids}} = \sum_i \sum_j r_{ij} d(x_i, m_j)$$

The k-Medoids algorithm exhibits superior robustness compared to k-Means due to employing actual data points as cluster centers. As medoids are chosen directly from the dataset, they are inherently more resistant to outliers and noise [1]. In contrast, k-Means can be heavily influenced by outliers, leading to less robust clustering results [3]. However, k-Medoids might not be adequate for huge datasets since it can be computationally expensive [11].

5.3 Cluster Analysis of Time Series Data: Identifying Similar Consumption Patterns

The objective of this section is to employ the k-Means clustering algorithm to identify and classify distinct groups of users in Barcelona based on their similar consumption patterns in time series data. In this analysis, the subset of of users with size 1000 containing a higher number of entries is selected, considering the presence of missing values in some of the user records.

According to the silhouette score analysis (see Section 5.5.1 for further details), which measures the cohesion and separation of data points within clusters [12], the number of clusters k maximizing the score was determined to be three. This suggests that the data is divided into three distinct groups or types of users based on their water consumption patterns.

5.3.1 Clustering Results and Interpretation of water consumption patterns

Figure 3 depicts the time series data, with each time series data colored according to the respective cluster assignment. This visualization enables the identification of distinct clusters based on the similarities in consumption patterns over time.

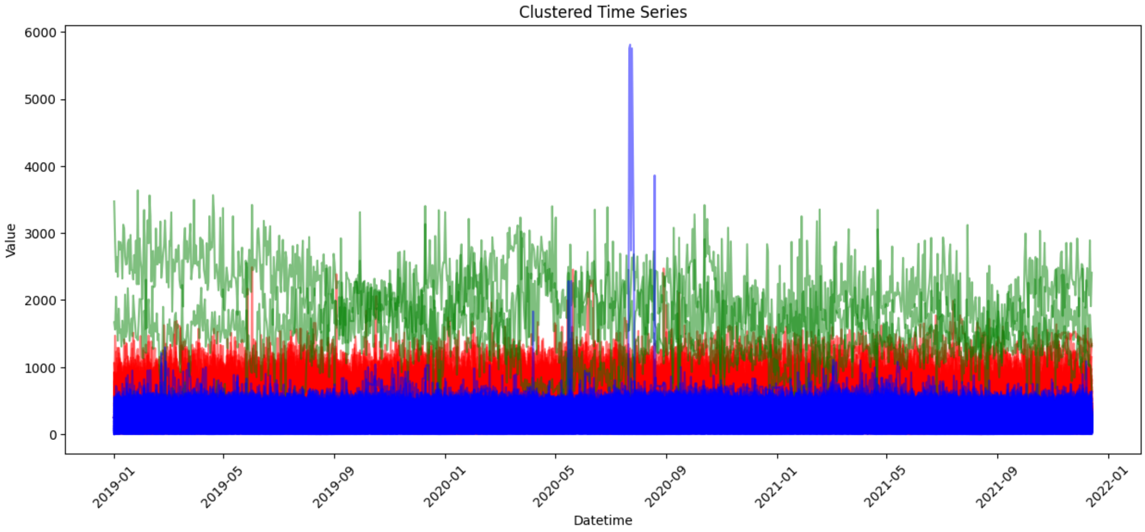
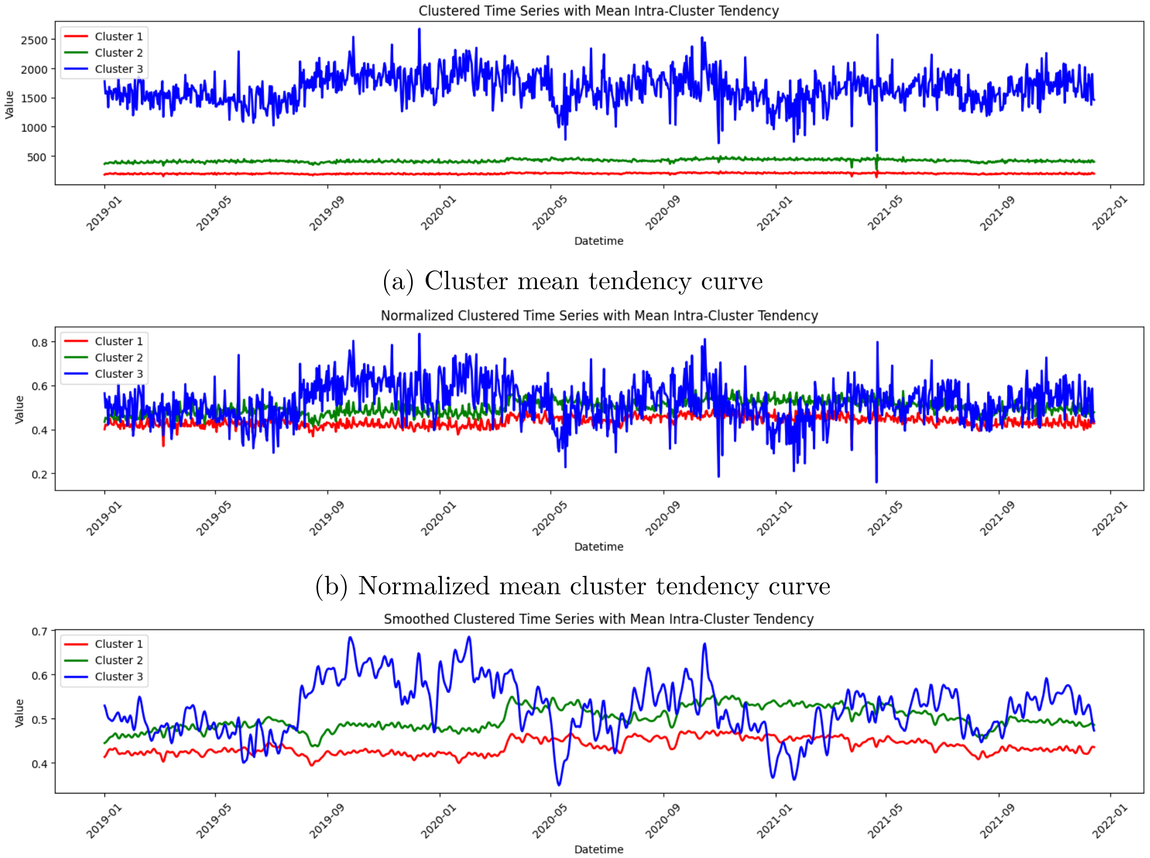


Figure 3: Water consumption time series colored by cluster

Figure 4a focuses on showing the mean tendency line of each cluster. The tendency line is derived by calculating the average consumption values for each time point within a cluster. This line serves as an informative summary of the collective behavior exhibited by the users in each cluster. Figure 4b showcases the normalized tendency lines obtained by applying

Min-Max scaling[9] to the data, thereby providing a fair comparison between cluster variations. Finally Figure 4c shows a smoothed version of tendency lines after applying a Gaussian filter to ease visualization.



(a) Cluster mean tendency curve

(b) Normalized mean cluster tendency curve

(c) Smoothed mean cluster tendency curve

Figure 4: Mean tendency for each cluster

Among these three clusters, one cluster stands out due to its relatively high water consumption values. Interestingly, this cluster exhibits a notable reduction in water consumption during the pandemic period. Higher water consumption may be associated with specific socioeconomic factors, such as higher income levels or different lifestyles. These factors could influence water usage patterns, for instance, through the presence of amenities like swimming pools or larger properties that require more water for maintenance. The reduction during the pandemic might be linked to economic impacts, lifestyle changes, or conscious efforts to conserve resources during COVID-19 crisis. In contrast, the other two clusters who experienced an increase during the pandemic may be associated to a low-medium socioeconomic class, where

[9] See Min-Max Scaling in Appendix A.4

individuals spending more time at home due to lockdown measures may have engaged in additional water consumption activities, such as increased cleaning, remote work from home, cooking, or gardening.

5.4 Identifying Clusters in Individual Water Usage

This chapter focuses on identifying the users with more distinguishable types of consumption patterns within their individual consumption by evaluating the cluster separation with the Calinski-Harabasz score, explained in Section 5.5.2. The motivation of this analysis is to identify excessive consumption behaviors and opportunities for improvement as well as helping to create personalized services and targeted interventions.

In order to ignore water meters with insufficient data points and obtain more accurate results, we have selected a subset of the dataset containing the 1000 water meters with more entries and computed the Calinski-Harabasz score for each of them. Then, the water meters have been sorted by descending score and the first 20 water meters were plotted. In this way, we could obtain a visualization of the most contrasted clusters.

Furthermore, evaluating the quality of k-Means clustering can also involve assessing other metrics, such as the silhouette score, proportion of variance explained[10], and Davies-Bouldin index [9]. These metrics provide additional insights into the quality of the clustering, such as the compactness and separation of the clusters [14]. Section 5.5 provides further details on clustering metrics.

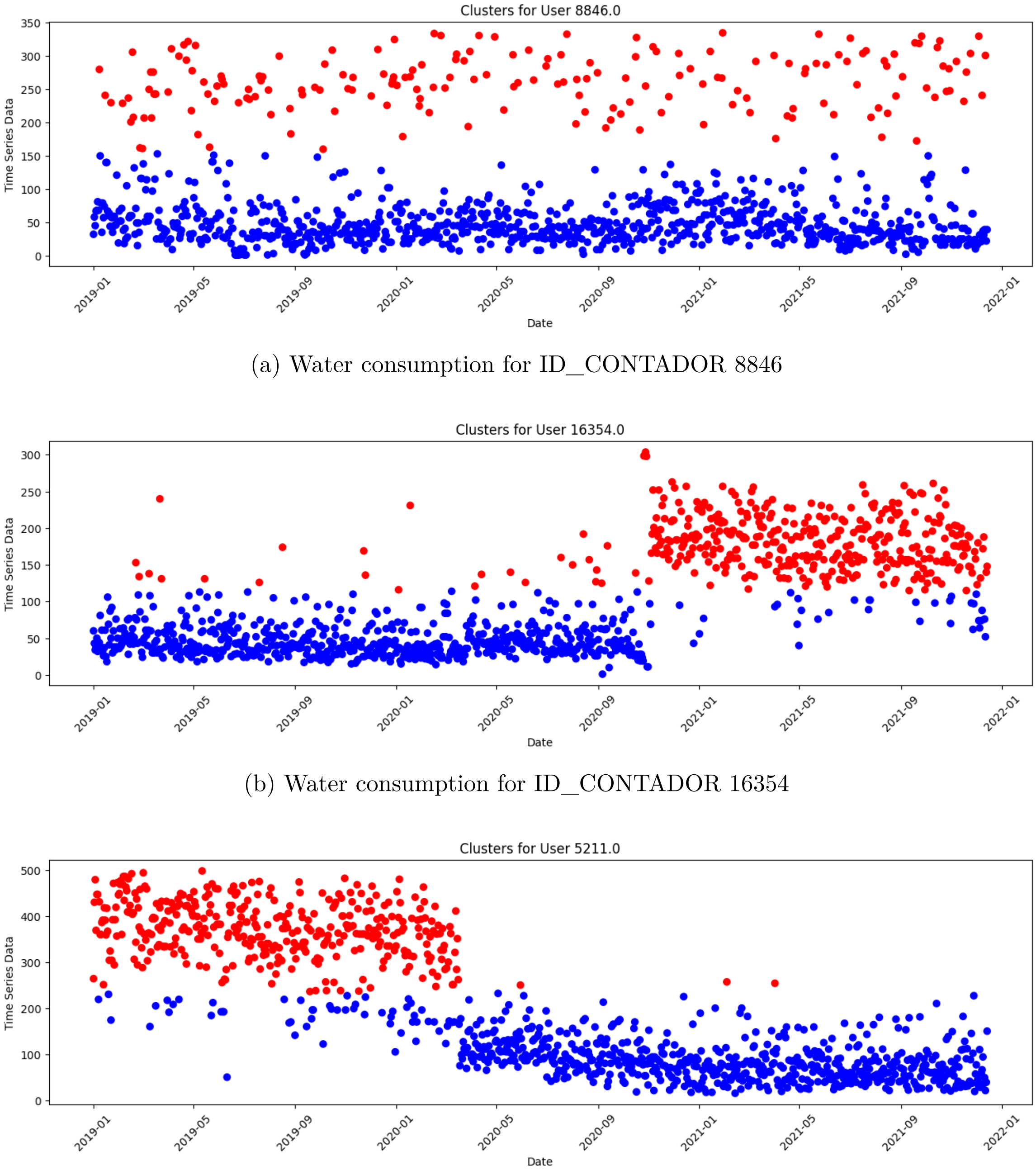
5.4.1 Clustering Results and Interpretation of Individual Water Usage

Figure 5a clearly shows two clusters with data points that oscillate in different ranges. This may be due to a clustering between weekdays and weekends, where the water consumption might significantly vary.

Figure 5b depicts the consumption time series data of a user who significantly increased their range of consumption. Possible hypotheses for this behavior are changes in household size, home renovations or upgrades or lifestyle changes.

In contrast to Figure 5b, Figure 5c shows a significant decrease in consumption. This variation may be due to a reduction in household size or lifestyle changes adopting a more sustainable behavior.

[10] See Proportion of Variance Explained in Appendix A.5



(a) Water consumption for ID_CONTADOR 8846

(b) Water consumption for ID_CONTADOR 16354

(c) Water consumption for ID_CONTADOR 5211

Figure 5: Water consumption for different ID_CONTADOR

5.5 Clustering Evaluation Metrics

5.5.1 Silhouette score

The silhouette score is a measure of how well a data point fits into its assigned cluster compared to other clusters. It ranges from -1 to 1, where a score of 1 indicates that the data point is well-matched to its own cluster and poorly matched to neighboring clusters, while a score of -1 indicates the opposite. A score close to 0 indicates that the data point is on or very close to the boundary between the two clusters [12].

Let $a(i)$ be the average distance between the data point and all other data points in the same cluster, which is called the average intra-cluster distance. Let $b(i)$ be the average distance between the data point and all data points in the nearest neighboring cluster, which is called the average inter-cluster distance. The silhouette score for each data point is then calculated as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

The overall silhouette score for a clustering system is computed as the average of the silhouette scores for all the data points in the dataset [12].

5.5.2 Calinski-Harabasz index

The Calinski-Harabasz index is a measure of how well-separated the clusters are from each other. It is calculated as the ratio of the between-cluster variance to the within-cluster variance, multiplied by the ratio of the number of data points minus the number of clusters to the number of clusters minus one [13]. The Calinski-Harabasz index, denoted as $CH(K)$ is computed as follows:

$$CH(K) = \frac{B(K)}{W(K)} \times \frac{N - K}{K - 1}$$

where K is the number of clusters, N IS the total number of data points, $B(K)$ is the between-cluster variance and $W(K)$ is the within-cluster variance. Between-cluster variance measures the separation between clusters using the distances between centroids, while within-cluster variance measures the compactness of the cluster using the distances of the cluster points to its corresponding centroid. Let n_k be the number of points in cluster k, c_k be the centroid of cluster k, c be the centroid of the dataset and x_{ik} be the i^{th} point of cluster k.

$$B(K) = \sum_{k=1}^{K} n_k \times ||c_k - c||^2$$

$$W(K) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} ||x_{ik} - c_k||^2$$

Higher values of the index indicate better-defined clusters [13].

5.5.3 Davies-Bouldin index

The Davies-Bouldin index is another measure of cluster separation that takes into account both the within-cluster and inter-cluster distances. It is calculated as the average similarity between each cluster and its most similar cluster. [14].

Let k be the number of clusters, $R(i, j)$ be the similarity between clusters i and j.

The Davies-Bouldin index is computed as:

$$\mathrm{DB} = \frac{1}{k} \sum \max\left(R(i, j)\right)$$

The similarity between clusters i and j is defined as:

$$R(i, j) = \frac{W(i) + W(j)}{B(i, j)}$$

where $W(i)$ and $W(j)$ represent the within-cluster variance, i.e, the mean sum of squared distances to the centroid of clusters i and j respectively. $B(C_i, C_j)$ represents the between-cluster variance, i.e, the sum of squared distances between centroids of clusters C_i and C_j. Higher scores indicate greater similarity and lower values of Davies-Bouldin index indicate better-defined clusters [14].

5.5.4 Evaluation results

Figures 5a, 5b, and 5c demonstrate relatively good clustering results with high silhouette scores, high Calinski-Harabasz indices, and low Davies-Bouldin indices. On the other hand, Figure 4 has lower evaluation scores, suggesting that the clusters in this plot might not be as well-defined or separated.

Table 1: Clustering Evaluation Metrics

Figure	Silhouette Score	Calinski-Harabasz	Davies-Bouldin
4	0.344	371	1.073
5a	0.825	6971	0.263
5b	0.749	5568	0.354
5c	0.766	6848	0.311

6 Scaling and normalization methods

Several inconsistencies were observed in the magnitude of the provided data, needing preprocessing techniques such as scaling. These inconsistencies could introduce bias or undue influence from variables with larger scales during analysis and modeling processes. To address this issue, scaling techniques such as z-score scaling and factor scaling were employed.

6.1 z-score Scaling for Standardization

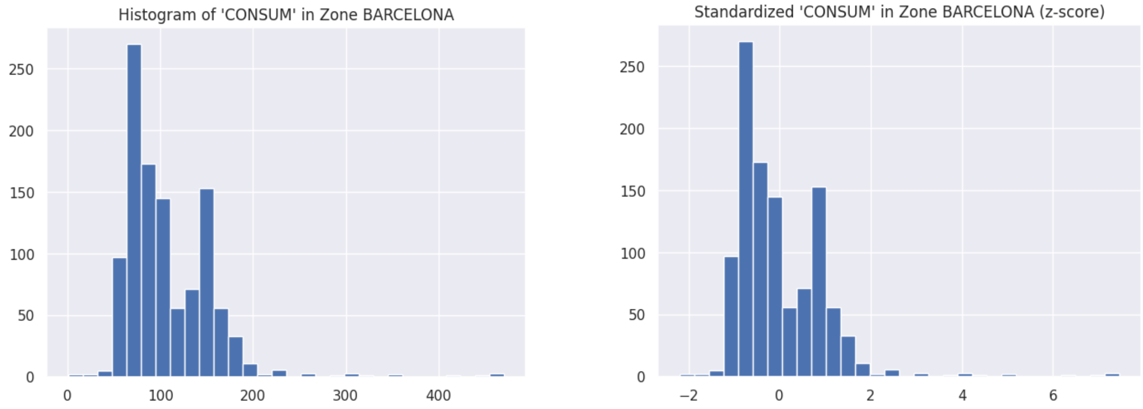
The z-score scaling, also known as standardization, is a data preprocessing technique that transforms data to have a mean of zero and a standard deviation of one without altering the underlying distribution [17]. This is done by subtracting the mean from each data point and then dividing by the standard deviation. The resulting transformed values are called z-scores.

The z-score scaling is useful when comparing variables that are measured on different scales or have different units of measurement [3]. By standardizing the data, it allows for direct comparison of the relative size and direction of the values of different variables. In addition, the z-score scaling is also used in statistical analysis as it enables the calculation of probability distributions and hypothesis testing [3]. Additionally, Stevens [18] suggested the use of standardization to facilitate comparison between different variables in psychological research. Figure 6b provides a visual representation of the standardized data.

6.2 Scaling factor

The scaling factor ξ is calculated as the ratio of the mean consumption before the specific date and the mean consumption after the specific date. This scaling factor is then used to scale the consumption data after the specific date, such that the mean consumption before and after the specific date is equal. Mathematically, the scaling factor ξ can be expressed as:

$$\xi = \frac{\frac{1}{n}\sum_{i<j} x(i)}{\frac{1}{m}\sum_{i>j} x(i)}$$



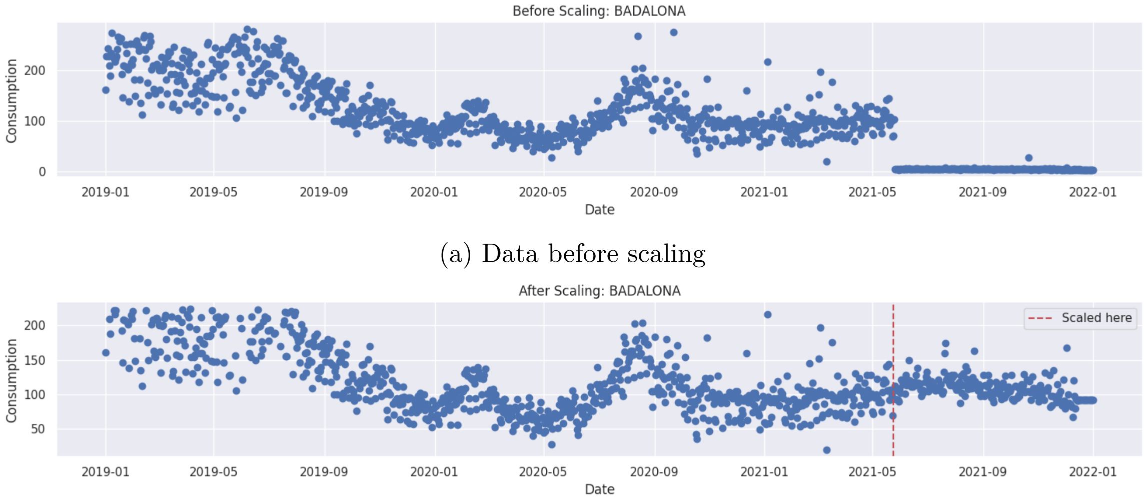
(a) Original data

(b) Standardized data

Figure 6: Histogram of original and standardized data

where $x(i)$ represents the consumption value at time i, j is the time from which the data has a wrong scale, n is the number of data points before j, and m is the number of data points after j.

Scaling the data in this way allows for better comparability of the consumption patterns before and after the specific date, and ensures that any observed differences are not only due to a change in consumption levels. Additionally, it can help to identify and adjust for outliers in the data that may affect the accuracy of any subsequent analysis. Figure 7b shows a graphical representation of the scaling.



(a) Data before scaling

(b) Data after scaling

Figure 7: Data before and after scaling

6.3 Shapiro-test statistic for normality

The Shapiro-Wilk test is a common statistical test for normality that compares the observed distribution of data to the expected normal distribution of data by computing the degree of similarity between these two. The goal order to determine if the data is likely to have a normal distribution [15].

According to Royston, the Shapiro-Wilk test is a more accurate approach than other tests for normality, such as the Kolmogorov-Smirnov test [15], for small to medium-sized samples [16]. The Shapiro-Wilk test is designed to identify deviations from normality on heavy tail or skewed data by using a weighting factor that gives more weight to data towards the tails of the distribution, which is where deviations from normality are most likely to occur [16].

However, note that the Shapiro-Wilk test has certain limitations. For instance, the test can be affected by the sample size, and it may fail to detect non-normality in large sample sizes [17]. In addition, the test assumes that the observations are independent and identically distributed [16], which is not the case of this application. Water consumption is influenced by external factors that are not constant or random, such as temperature, rainfall, holidays, and other events. Thus, the patterns observed in one period may not necessarily hold in another, making the data dependent on recent observations and requiring conditional probabilities.

Therefore, the non-iid nature of the data of the AB Data Challenge poses a challenge for data analysis and requires careful consideration and appropriate statistical techniques to account for the dependencies in the data.

7 Polynomial Regression Curve Fitting

Polynomial regression is a technique used to find the best polynomial function that fits a set of data. It consists of estimating the coefficients of a polynomial equation of degree n that can fit the data points with minimum error [19]. The equation takes the general form:

$$P(x) = b_0 + b_1x + b_2x^2 + \ldots + b_nx^n$$

where $P(x)$ is the dependent variable, x is the independent variable, and $b_0, b_1, b_2, \ldots, b_n$ are the coefficients of the polynomial equation.

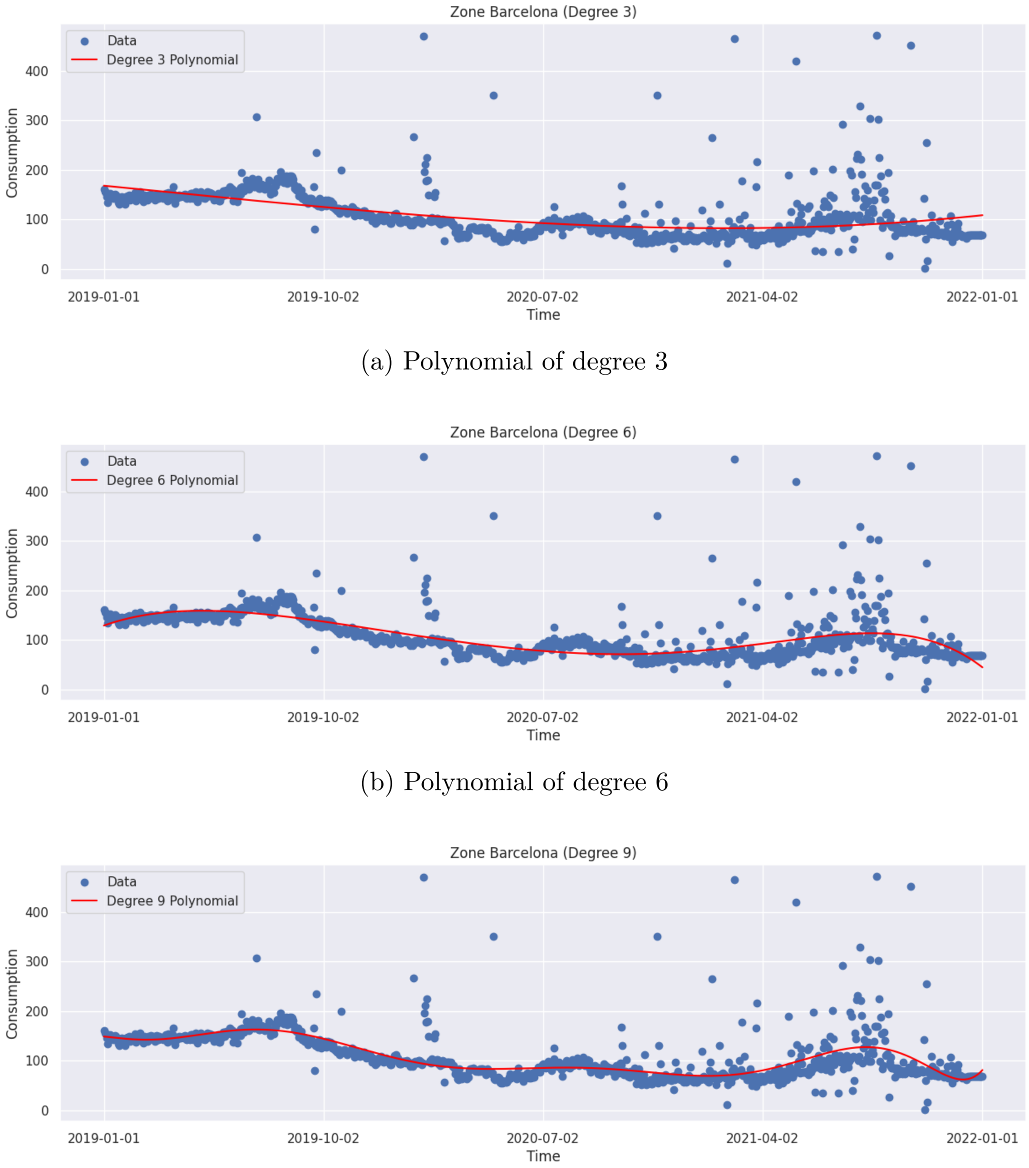
According to Draper, N. R., & Smith, H. , to compute the best polynomial fit, we need to determine the values of the coefficients $b_0, b_1, b_2, \ldots, b_n$ that minimize the sum of squared errors between the predicted values of the polynomial and the actual values in the dataset. This can be achieved using a method called least-squares regression. [19]. In this method, the coefficients are estimated by solving a system of linear equations based on the data points, using matrix algebra[11]. The function to be minimized is denoted as:

[11] See solving the linear system in Appendix A.6

$$SSE = (y_1 - P(x_1))^2 + (y_2 - P(x_2))^2 + \ldots + (y_n - P(x_n))^2$$

where $y_1, y_2, \ldots, y_n$ are the observed values and $P(x_1), P(x_2), \ldots, P(x_n)$ are the estimated values.

As the degree of the polynomial is increased, the model better matches the training data, but it can also result in overfitting, when the model becomes too complex and fails to translate well to new data. As a result, a balance must be achieved between the model's complexity and its capacity to fit the data without overfitting. [19]. Figure 8a, Figure 8b and Figure 8c show the estimated polynomial for different degrees.



(a) Polynomial of degree 3

(b) Polynomial of degree 6

(c) Polynomial of degree 9

Figure 8: Polynomials of different degrees

8 Spatial Analysis of the Barcelona Metropolitan Area

Figure 9 shows the median consumption for municipalities in the metropolitan area of Barcelona from 2019 to 2021. Visual analysis suggests that cities in the surroundings of Barcelona, often with less population, have higher water consumption. For example, Badalona has a population of over 200,000 but a relatively low domestic water consumption per person per day of 102.36 liters, while the smaller municipality of Tiana has a much higher consumption rate of 147.3 liters.

Tiana has the highest consumption rate of 147.3 liters per person per day, while Badia has the lowest consumption rate of 86.42 liters per person per day.

Overall, the average median domestic water consumption for the municipalities in the Barcelona metropolitan area is around 107 liters per person per day. However, there is a lot of variation within this average, indicating that factors such as location and infrastructure can have a significant impact on water consumption patterns.

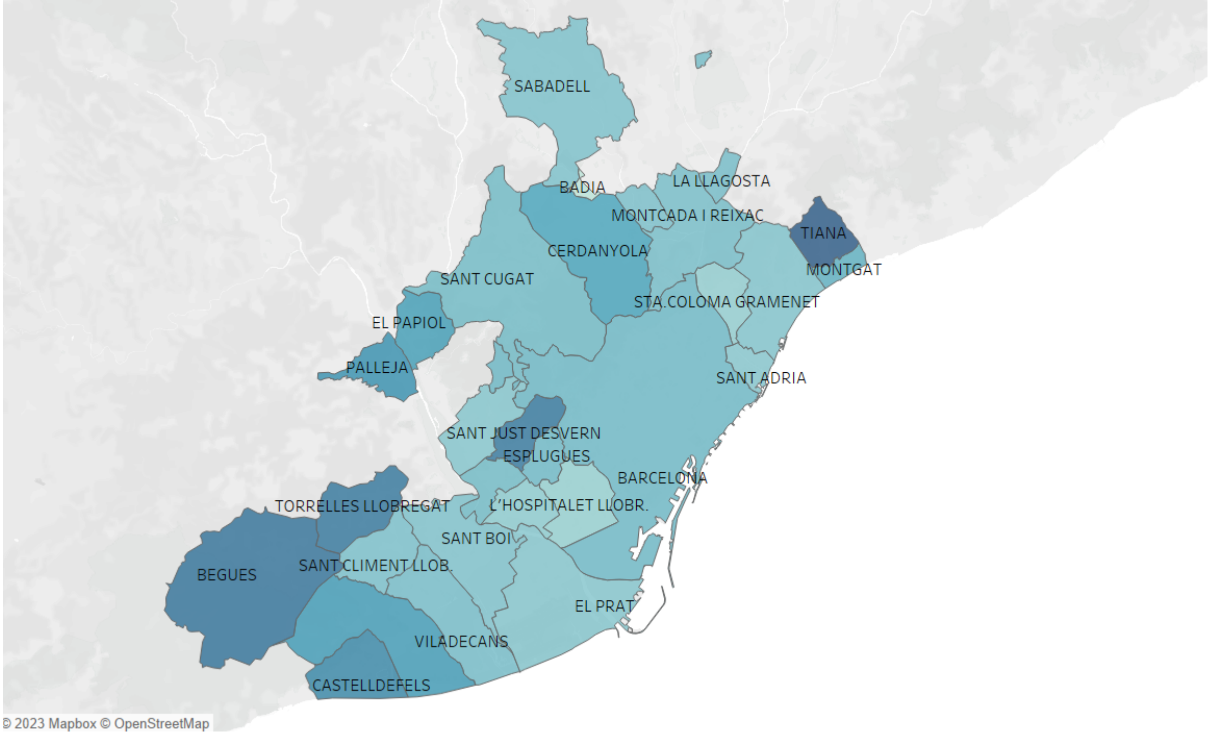


Figure 9: Median domestic consumption in the Barcelona Metropolitan Area

9 Impact of the COVID-19 Pandemic on Water Consumption in the Municipalities of the Metropolitan Area of Barcelona

The pandemic of COVID-19 has had a significant effect on several aspects of daily life, including domestic water consumption. As individuals were encouraged to wash their hands frequently and follow additional hygiene measures, the demand for domestic water has changed noticeably. Water usage habits have been modified by lockdown measures, work-from-home, and altered routines, resulting in both increases and shifts in domestic water consumption. The zones with the highest percentage increase[12] in domestic water consumption, such as La Llagosta (212.86%), Sabadell (100.85%), and Montgat (45.87%), could be areas with relatively high population density or a large number of multi-person households, which could explain the significant increase in water consumption. There is no clear correlation with the COVID-19 cases[13]. On the other hand, the zones with the lowest percentage increase or a decrease in domestic water consumption, such as Santa Coloma de Cervelló (-69.84%), Viladecans (-42.34%), and Sant Climent de Llobregat (-37.22%), could be areas with a lower population density or areas with a higher number of single-person households, which could result in an overall decrease in water consumption.

The significant decrease in water consumption in zones such as Santa Coloma de Cervelló and Viladecans could be due to a variety of factors, such as the closure of businesses and industries that consume large amounts of water or the migration of residents away from these areas due to the pandemic. Additionally, these areas presented a higher consumption compared to the other municipalities and may experienced a reduction due to the factors discussed in Section 5.4.1.

10 Factors Influencing Water Consumption

This chapter aims to investigate the various factors that influence water consumption patterns. The following variables will be examined in relation to their impact on water usage: economical resources, temperature, precipitation, relative humidity, solar radiation, population density, weekday, and seasonal variations.

10.1 Multiple Linear Regression and SHAP Analysis

Multiple linear regression provides interpretable coefficients, allowing us to understand the relationship between each input variable and the target variable [10]. Multiple linear regression assumes a linear relationship between the input variables and the target variable [19]. While this assumption may not hold perfectly in practice, it often provides a good approximation for many real-world scenarios. Multiple linear regression is a computationally efficient model, especially with a relatively small number of input variables [10]. It can handle a moderate number of features without significant computational overhead [3].

SHAP (SHapley Additive exPlanations) values, originally derived from cooperative game theory, have been adapted and applied in the context of machine learning interpretability.. SHAP values provide a quantitative measure of the impact of each feature on the model's output, allowing for the interpretation of feature importance [31]. Positive SHAP values indicate a positive contribution of a feature to the prediction, while negative values suggest a negative contribution [30]. The magnitude of the SHAP value reflects the relative impor-

[12] Considering the lockdown period from March 14th to June 21th of year 2020

[13] Dades Obertes de la Generalitat. {https://analisi.transparenciacatalunya.cat/Salut/Registre-de-casos-de-COVID-19-a-Catalunya-per-muni/jj6z-iyrp

tance or influence of the feature on the model's prediction [31]. Furthermore, the sum of SHAP values for a prediction equals the difference between the model's prediction for that instance and the expected average prediction, ensuring local accuracy and consistency in the interpretation of feature importance across predictions [30]. Thus, by analyzing SHAP values, researchers can identify important features, discern intricate relationships between variables, and enhance transparency and trust in machine learning models [31]. Figure 10 shows a SHAP summary plot.

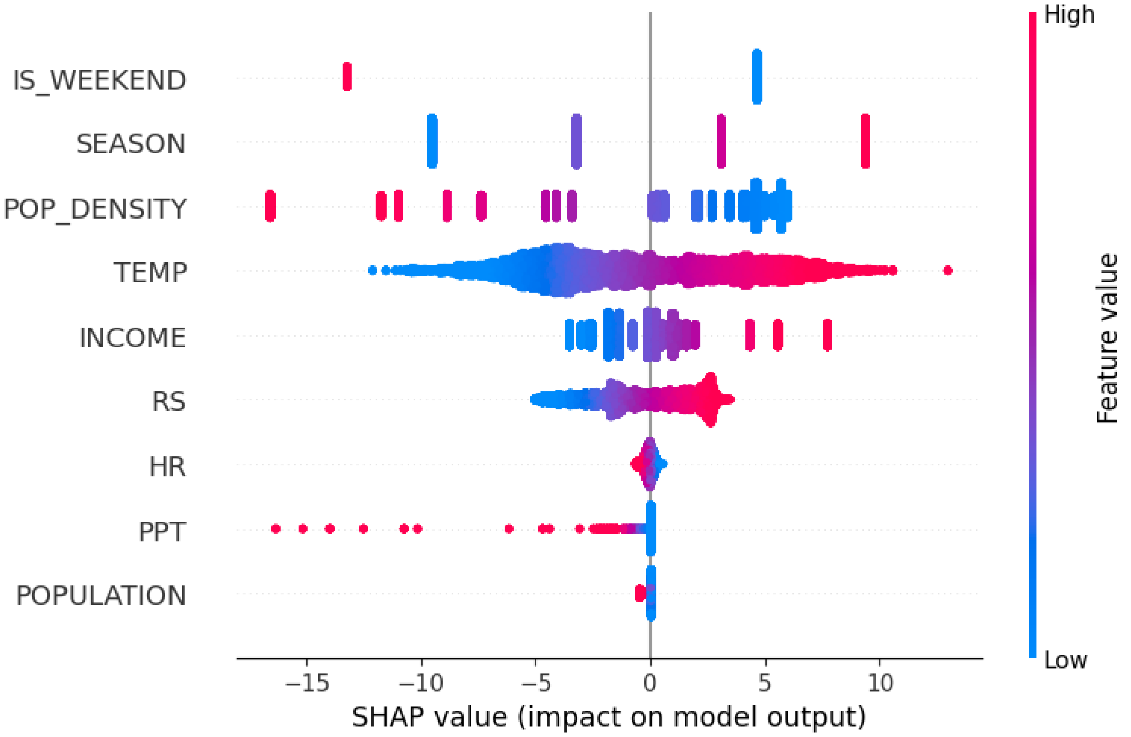


Figure 10: Summary plot of SHAP Values for targeted variables

The following variables are sorted decreasingly according to the contribution to the water consumption:

- Weekday: Counterintuitively, weekdays tend to have higher water consumption due to increased demand for activities like laundry, dish washing, and personal hygiene.
- Season of the year: During warmer seasons, such as summer, water usage tends to increase due to greater outdoor activities and increased demand for irrigation. The next section shows a butterfly plot to confirm the seasonal pattern.
- Population density: Higher population densities generally result in lower water demand. This is likely due to the fact that municipalities with lower population density are associated with higher income.

- Temperature: During hotter periods, people tend to use more water for activities like showering, watering gardens, and using air conditioning, which can increase overall water consumption.
- Gross annual household income: Individuals with higher economical resources might live in houses with swimming pools or installations that can increase the water consumption
- Solar Radiation: In areas with high solar radiation, there may be increased demand for outdoor irrigation and watering of plants.
- Relative humidity: Higher relative humidity can impact water consumption, particularly with regard to indoor water use. In more humid conditions, individuals may use less water for activities like humidifiers, air conditioning, or watering indoor plants.
- Precipitation: In regions with high rainfall, there might be less need for outdoor irrigation, resulting in lower water usage. Conversely, areas with low precipitation may rely more on artificial irrigation, leading to higher water consumption.
- Population: Number of inhabitants seems to have a minimal effect on water consumption, which is more affected by variables such as population density.

10.2 Consumption Patterns

By analyzing the violin plot shown in Figure 12, one can observe the relative heights and widths of the violins, indicating the differences in water consumption across seasons. If the summer violin is higher and wider compared to other seasons, it suggests that water consumption tends to be higher during the summer months. Conversely, if the winter violin is narrower and shorter, indicates lower water consumption during winter.

Figure 11 shows a decreasing trend from 2019 to 2021, especially in the years following the COVID-19 pandemic. This could be due to awareness and behavioral changes in uncertain times. Additionally, the plot highlights a seasonal pattern in water consumption, with higher levels during the warmer months.

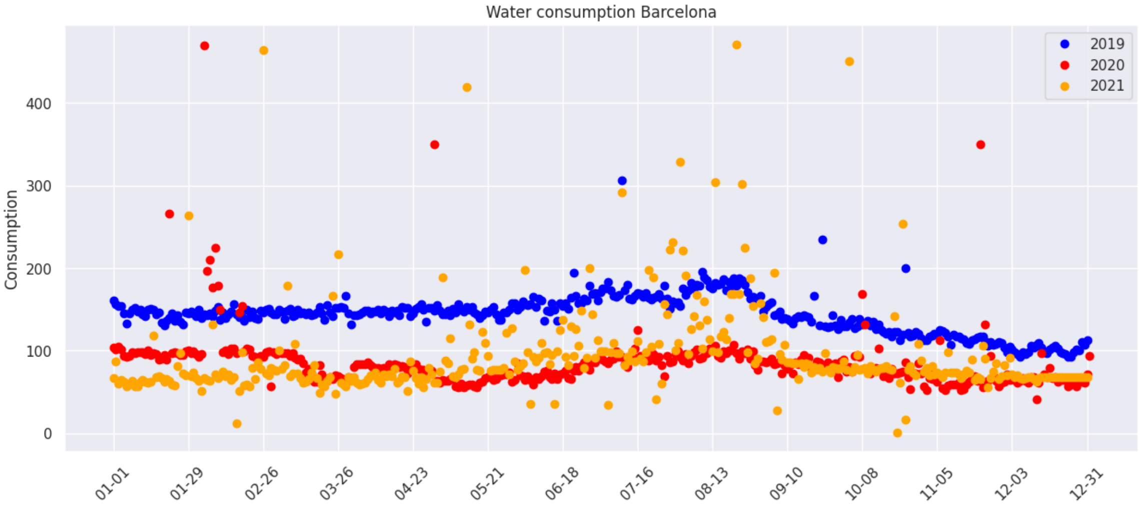


Figure 11: Water consumption in Barcelona from 2019 to 2021

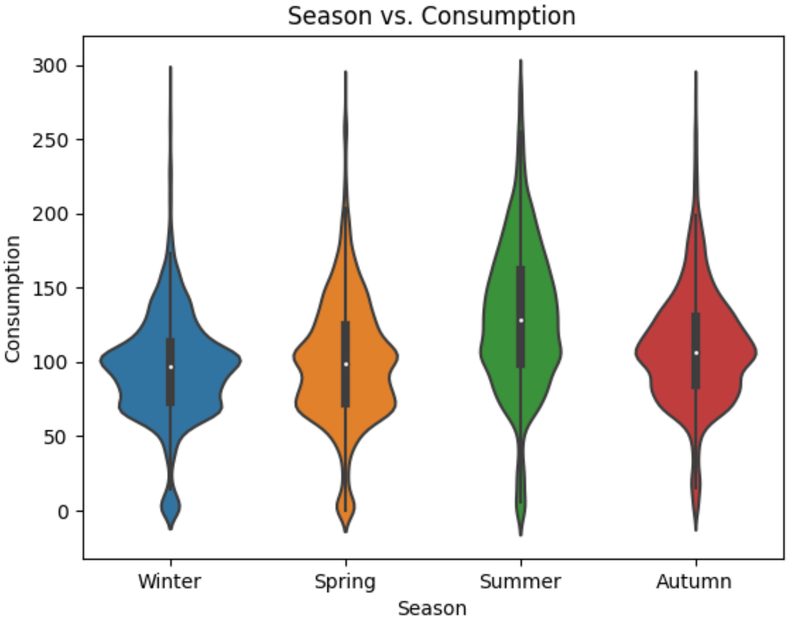


Figure 12: Violin plot showing seasonal variations

11 Time Series Forecasting using Autoregressive Integrated Moving Average Model

11.1 Justification of ARIMA model

The ARIMA (Autoregressive Integrated Moving Average) model is a suitable choice for forecasting water consumption due to its ability to capture trends, and past dependence in the data [22]. While water consumption data may not be strictly stationary, the ARIMA model can handle this by applying differencing to make the series approximately stationary. It provides a straightforward and interpretable framework, with model parameters having direct interpretations [28]. The ARIMA model's performance in resource consumption forecasting has been demonstrated in various studies such as the work from T. Jakaša, I. Andročec and P. Sprčić in 2011 [29], offering reliable and robust results. The computational efficiency and availability of software packages further facilitate the practical implementation of the model. The ARIMA model justifies its use for water consumption forecasting, offering a balance between accuracy, interpretability, and practicality.

11.2 Components of ARIMA model

The ARIMA model combines autoregressive (AR), differencing (I), and moving average (MA) components. It is characterized by three key parameters: p, d, and q [32].

1. Autoregressive (AR) model: Y_t is predicted by a linear combination of its previous values with some lag order. This is represented by the equation below, where c is a constant, ϕ is the magnitude of the autocorrelation, p is the number of lags, and ϵ_t is the error term [28] [32].

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \epsilon_t$$

2. Moving average (MA) model: Y_t is predicted as a linear combination of past error terms (ϵt). In the equation below, i θ is the value of the autocorrelation of the errors, and q is the number of lags [28] [32].

$$Y_t = c + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q}$$

3. Differencing (Integration): In an ARIMA model, the time series being modeled must be stationary to obtain meaningful predictions.The differencing component d specifies the order of differencing applied to the series before estimating and is used to make the time series stationary. Differencing refers to calculating the difference between adjacent observations [28] (Schaffer, A.L., Dobbins, T.A. & Pearson, SA, 2021).

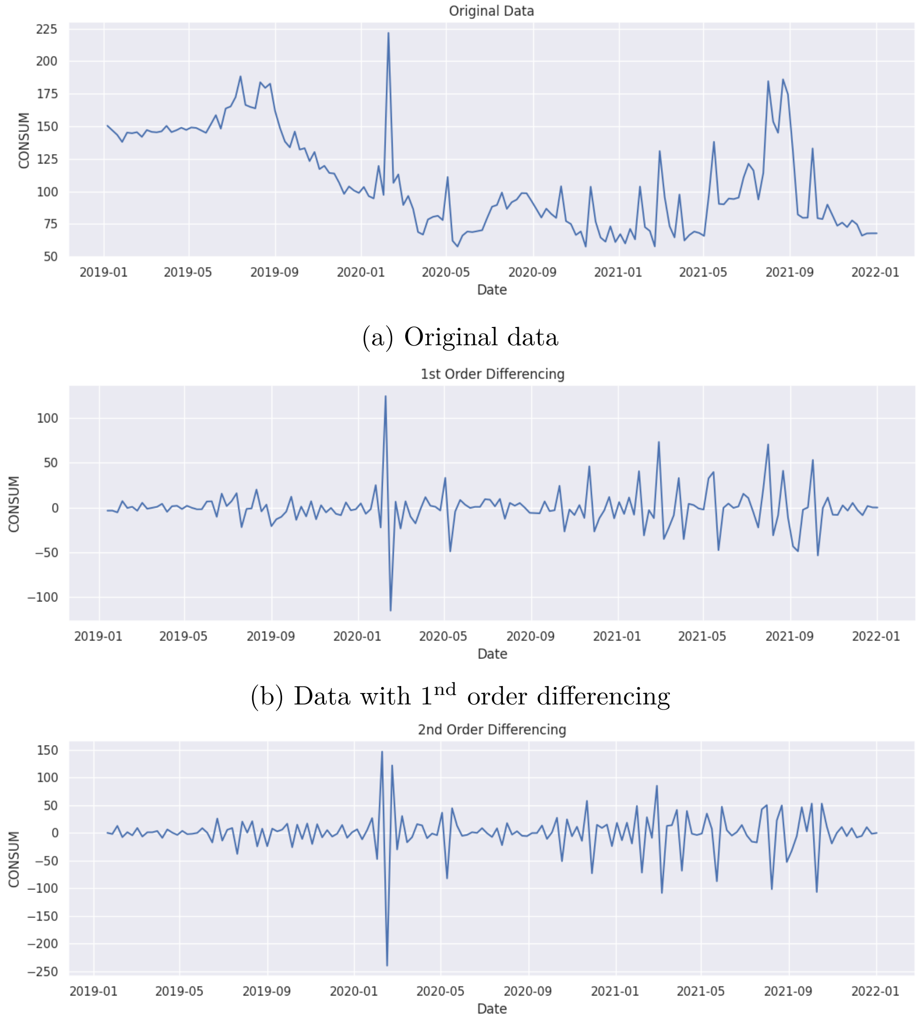
$$Y_t' = Y_t - Y_{t-1}$$

The ARIMA model is then built by differentiating the data at least once to make it stationary and combining the AR and the MA terms [28] [29]. So the equation of an ARIMA model becomes [28]:

$$Y_t = \alpha + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q}$$

11.3 Choice of parameters p, q, d

The first step would be to take care of the assumptions discussed above. For that, we need to determine the order of differencing d such that the data becomes stationary [29]. Figure 13 shows the data differenced until 2^{nd} order. Figure 13b shows that at 1^{st} order differencing, data is approximately stationary, so we can set the parameter d as 1.



(a) Original data

(b) Data with 1^{nd} order differencing

(c) Data with 2^{nd} order differencing

Figure 13: Differenced data until 2^{nd} order

The autocorrelation function (ACF) plot shows the correlation between the time series and its lagged values [22]. The blue area depicts the 95% confidence interval and is an indicator of the significance threshold. The cutoff for significant spikes is determined by observing when the values cross the confidence interval. Figure 14 shows the ACF plot with a significant correlation on lag 4, suggesting a potential autoregressive order q of 4.

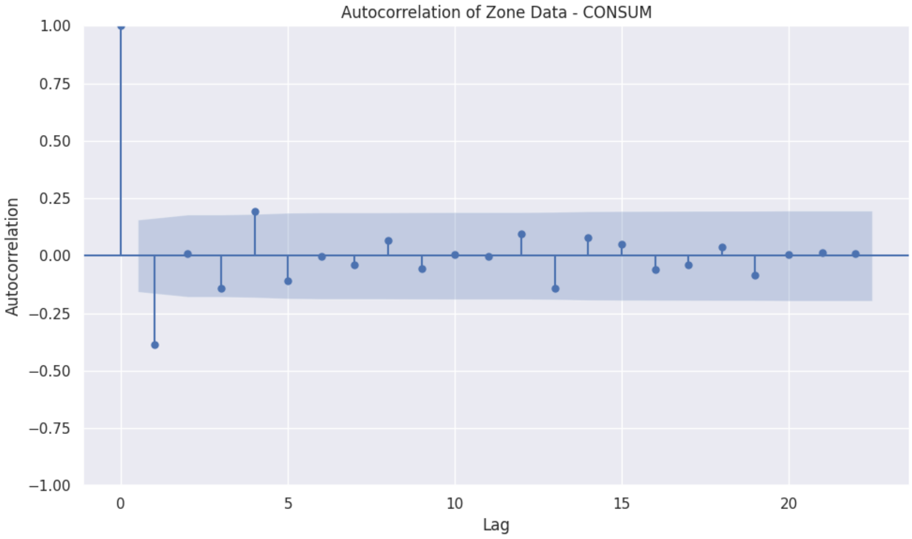


Figure 14: Autocorrelation plot

The partial autocorrelation function (PACF) plot, represents the correlation between the series and its lagged values after removing the effects of earlier lags. Figure 15 depicts the PACF plot where spikes indicate a potential autoregressive relationship up to order $p = 3$. Thus, the choice for the ARIMA model is ARIMA(p, d, q) = ARIMA(3, 1, 4).

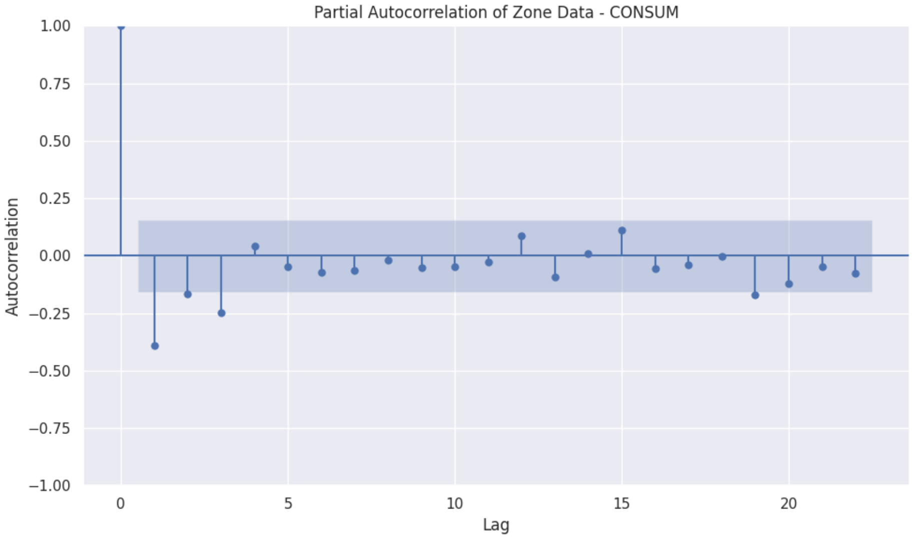


Figure 15: Partial Autocorrelation plot

It's important to note that the selection of ARIMA parameters is not solely based on the

PACF and ACF plots but should also consider other factors such as model diagnostics, AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) values [26].

11.4 Forecasting results of the ARIMA Model

Figure 16 shows a good agreement between predicted and observed values, indicating that the model has satisfactory predictive ability. The training data set includes consumption values from 2019 to 2021, while the testing data set includes values from 2021 to 2022. The model captures a trend of the data and predicts consumption peaks.

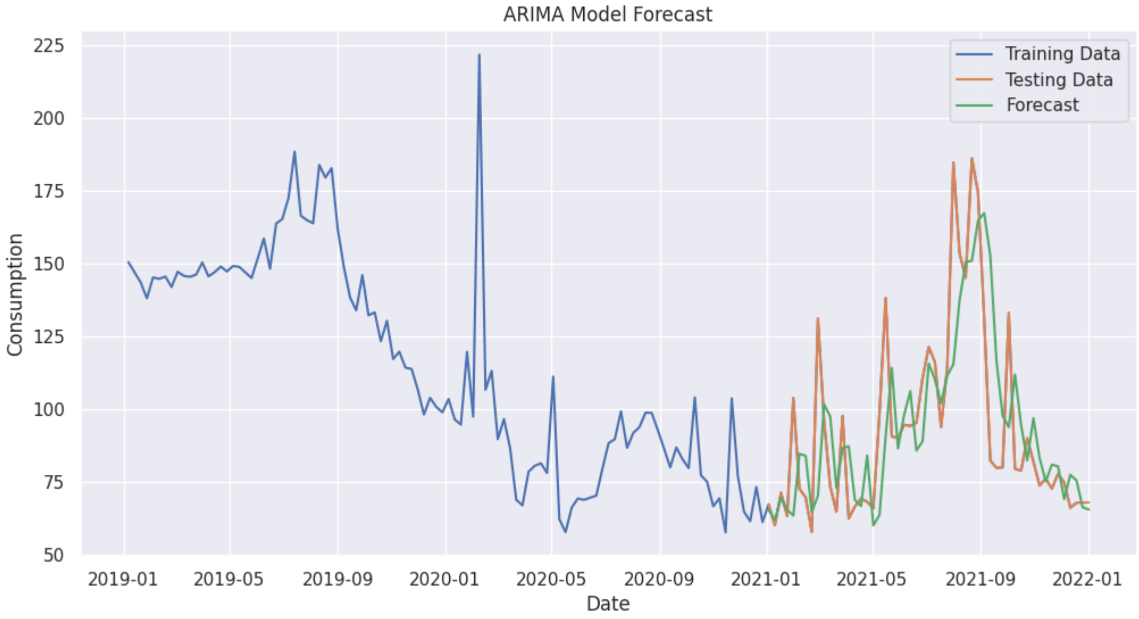


Figure 16: Observed and fit values for water consumption in Barcelona

Figure 17 depicts the forecasted values of ARIMA$(3, 1, 4)$ and shows a 95% confidence interval in a gray area.

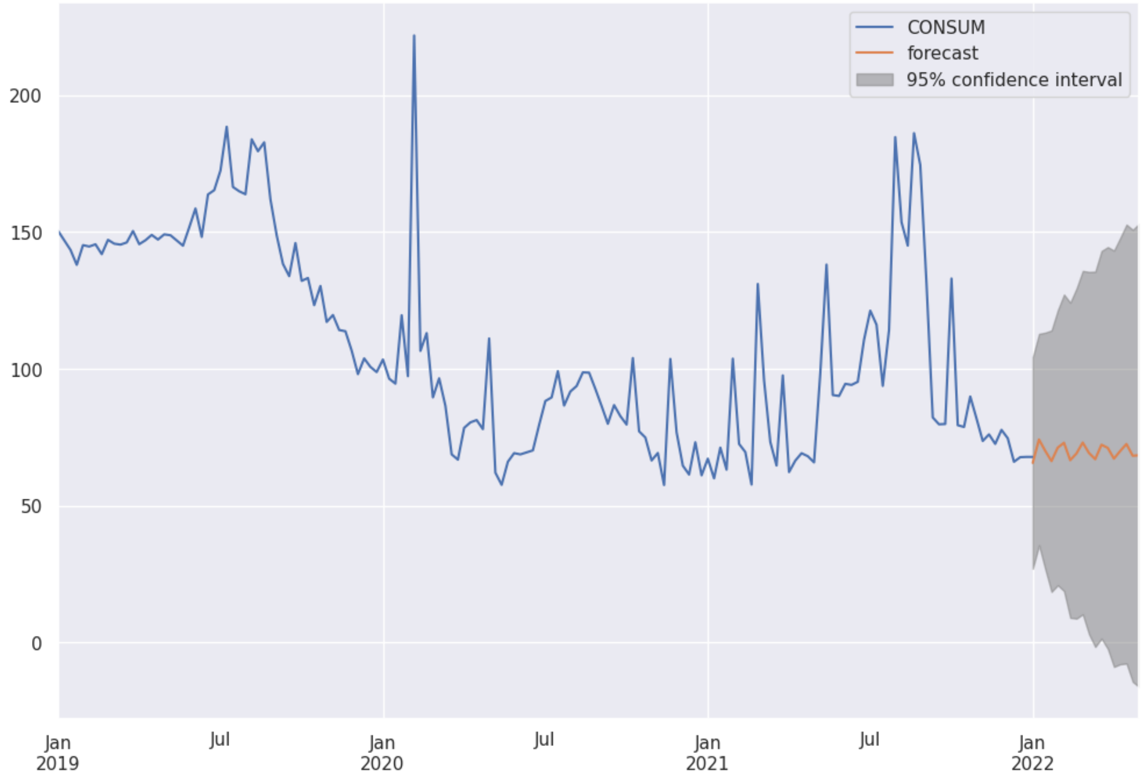


Figure 17: Forecast of water consumption in Barcelona

11.5 Evaluation metrics and performance of the ARIMA model

ARIMA(3,1,4) model demonstrates a moderate level of forecast accuracy, with a Mean Absolute Percentage Error (MAPE) of approximately 16.99% and a Mean Absolute Error (MAE) of approximately 16.58 units of consumption. The model shows a slight positive bias (ME), with a Root Mean Squared Error (RMSE) of approximately 23.89 units. It captures most of the autocorrelation in the data (ACF1 of approximately 0.012) and exhibits a moderately positive correlation (Corr) of approximately 0.68 between the forecasted and actual consumption values. The minimum absolute deviation (Min-Max Error) between the forecasted and actual values is approximately 0.94 units of consumption [21]. Overall, the ARIMA model performs reasonably well in forecasting consumption values. Table 2 shows the results for the corresponding metrics. Appendix A.7 shows the mathematical formulation for the used metrics.

Metric	Value
MAPE	16.998
ME	0.191
MAE	16.578
RMSE	23.887
ACF1	0.012
Corr	0.685
Min-Max err.	0.934

Table 2: Evaluation metrics for the ARIMA model

11.6 Comparative analysis with other Machine Learnings models

The ARIMA model demonstrated effective performance in forecasting values in this study. However, it has some limitations when capturing non-linear patterns, long-term forecasting, handling non-stationarity, and requiring manual parameter selection. Alternative machine learning models such as Long Short-Term Memory (LSTM), Prophet model, and Gradient Boosting Regression Trees (GBRT) showcased advantages in capturing complex relationships and handling non-linear data [34] [33] [32].

GBRT model supports both univariate and multivariate non-linear time series data but requires feature engineering and is prone to overfitting if not carefully tuned [27]. Prophet model incorporates seasonality, trends, and holiday effects and captures and multiple trends effectively. It is suitable for short to medium term forecasting tasks [33]. LTSM is based on a recurrent neural network (RNN) which captures complex temporal dependencies and long-term patterns while automatically learns relevant features from the data [34]. However, it requires a large amount of training data to perform optimally [33]. In general, these models may require additional computational resources, extensive hyperparameter tuning, and more sophisticated feature engineering.

Using advanced extensions of the ARIMA model such as SARIMA (Seasonal ARIMA) or SARIMAX (ARIMA with exogenous variables) can address the limitations of the standard ARIMA model [34]. SARIMA model includes a seasonality component and provides better performance for data with distinct seasonal trends [28]. However, it is computationally expensive compared to ARIMA [29]. The model ARIMAX extends the ARIMA model by incorporating exogenous variables, which are external factors that can influence the time series. It is suitable when there are known factors that influence the time series and can be measured or predicted but is complex to estimate and tune due to the inclusion of additional predictors [34]. Furthermore, automated parameter selection algorithms, such as grid search or automated ARIMA (auto-ARIMA), can assist in finding optimal model configurations [33]. Appendix A.8 provides further details with a brief overview of the mentioned models.

12 Conclusions

The project's findings and methodology offer a solid foundation for future research and practical implementations in the field of water management, providing opportunities for extending improvements and improvements in ensuring water sustainability in Barcelona and beyond.

The application of k-Means clustering provides a quantitative framework for objectively examining consumption similarities and differences, enabling data-driven insights and evidence-based decision-making. This method has the potential to uncover hidden consumption patterns and identify anomalies or outliers, which can serve as indicators of potential inefficiencies or areas requiring further investigation.

Users with lower consumption levels appear to be more sensitive to changes in their water usage habits during the pandemic. The increase in their water consumption during the COVID-19 crisis may be attributed to various factors, such as spending more time at home and engaging in additional household activities.

Many conclusions can be drawn based on factors identified to influence water use in Barcelona, notably weekdays, population density, season, temperature, gross yearly income, and solar radiation. To begin, the analysis demonstrates the importance of taking into account temporal elements such as weekdays and seasonality in water management. Weekday water use is expected to be greater due to increased residential and commercial activity. Furthermore, seasonality plays a crucial role, with higher consumption during warmer months.

This knowledge allows Aigües de Barcelona to anticipate and plan for fluctuations in demand, ensuring adequate supply and optimizing resource allocation. Some of the measures that can be taken during peak seasons include water restrictions, rainwater harvesting and water reuse systems to mitigate shortages.

Geospatial analysis and correlation between lower population density and higher water consumption suggest that factors such as larger residential properties, increased outdoor water usage, and higher socio-economic status may contribute to this pattern. Municipalities with low population density in the surrounding area of Barcelona count with Baix Llobregat, Maresme and Garraf aquifers. Exploring the feasibility of extracting water from these aquifers can provide a sustainable water source to meet the increasing demand in these areas.

The ARIMA model has shown robust performance in capturing and modelling the complex dynamics of water consumption, enabling accurate forecasting of future consumption trends. Evaluating the model using appropriate metrics such as mean absolute error (MAE) or root mean square error (RMSE), has indicated its ability to generate accurate predictions. However, the model accuracy would be improved by including a seasonality component and taking into account the variables affecting water consumption.

A Annex

A.1 Schema of AB Data

```
new
  DOMESTIC
    BADALONA.csv
    BARCELONA.csv
    BEGUES.csv
    CASTELLDEFELS.csv
    CERDANYOLA.csv
    CORNELLA.csv
    EL PAPIOL.csv
    ESPLUGUES.csv
    GAVA.csv
    LES BOTIGUES SITGES.csv
    L_HOSPITALET LLOBR.csv
    MONTCADA I REIXAC.csv
    MONTGAT.csv
    PALLEJA.csv
    SANT ADRIA.csv
    SANT BOI.csv
    SANT CLIMENT LLOB..csv
    SANT FELIU LL..csv
    SANT JOAN DESPI.csv
    SANT JUST DESVERN.csv
    STA.COLOMA CERVELLO.csv
    STA.COLOMA GRAMENET.csv
    TORRELLES LLOBREGAT.csv
    VILADECANS.csv
  Activitat-per-zona.zip
  COMERCIAL.zip
  ConsumPerCP.csv
  ConsumPerDistricte.csv
  ConsumPerSeccioCensal_.zip
  ConsumPerZona-1.csv
  Domestic_1.zip
  Domestic_2.zip
  Domestic_3.zip
  Domestic_4.zip
  INDUSTRIAL.zip
```

A.2 Structure of CSV files

Table 3: Sample records from BARCELONA.csv

	FECHA (date)	USO (use)	CONSUMO (consumption)	ID_CONTADOR (counter id)
10126881	2021-11-06	DOMÈSTIC	2219	17469
10126882	2021-11-02	DOMÈSTIC	74	8841
10126883	2021-11-07	DOMÈSTIC	27	8841
10126884	2021-11-07	DOMÈSTIC	206	18528
10126885	2021-11-03	DOMÈSTIC	80	18528

Table 4: Sample records from ConsumPerZona-1.csv

	FECHA	ZONA	CONSUM
29171	2021-12-14	TIANA	7430
29172	2021-12-14	BEGUES	397
29173	2021-12-14	STA.COLOMA GRAMENET	627
29174	2021-12-14	L'HOSPITALET LLOBR.	584
29175	2021-12-14	VILADECANS	483

A.3 Variables and magnitudes

Consumption	Litres (L)
Population density	Number of habitants/km^2
Gross annual household income	Euros (€)
Solar Irradiation	Watts per meter squared (W/m^2)
Temperature	Celsius degrees (°C)
Relative Humidity	%
Precipitation	millimetres (mm)

A.4 Min-Max Scaling

Min-max scaling, is a data preprocessing technique used to transform numerical data into a specific range, typically between 0 and 1. It involves linearly rescaling the original values based on the minimum and maximum values of the dataset. The formula for min-max scaling is as follows:

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

where x represents an original data point, $min(x)$ represents the minimum value of the dataset, $max(x)$ represents the maximum value of the dataset, and x' represents the scaled value.

A.5 Proportion of Variance Explained

Proportion of Variance Explained (PVE) measures the proportion of the total variance in the data that is accounted by the clustering. A higher percentage of variance explained indicates that the clustering has effectively captured the underlying patterns in the data. Several studies have used this approach to evaluate the performance of the k-means algorithm, such as in the work of Jain and Dubes [8].

PVE was computed with the following steps:

Let n be the total number of data points, x_i be the i-th data point, $\bar{x}$ be the mean of all the data points, and c_i be the centroid given to the i-th data point.

The total sum of squares (TSS) can be computed as:

$$TSS = \sum((x_i - \bar{x})^2) \quad \text{for } i = 1 \text{ to } n.$$

The sum of squared errors (SSE) can be computed as:

$$SSE = \sum((x_i - c_i)^2) \quad \text{for } i = 1 \text{ to } n.$$

The proportion of variance explained (PVE) can be computed as:

$$PVE = 1 - \frac{SSE}{TSS}.$$

A.6 Minimizing Least Squares Error

The system can then be solved by multiplying both sides by the transpose of the matrix of predictor variables and taking the inverse of the resulting matrix:

$$\begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^n \\ 1 & x_2 & x_2^2 & \dots & x_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & x_m^2 & \dots & x_m^n \end{bmatrix}^T \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^n \\ 1 & x_2 & x_2^2 & \dots & x_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & x_m^2 & \dots & x_m^n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^n \\ 1 & x_2 & x_2^2 & \dots & x_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & x_m^2 & \dots & x_m^n \end{bmatrix}^T \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

A.7 Mathematical formulation for clustering evaluation metrics

- MAPE (Mean Absolute Percentage Error):

  $$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$$

  where: y_i represents the observed value at time i, $\hat{y}_i$ represents the predicted value at time i, n represents the total number of observations [21].

- ME (Mean Error):

  $$\text{ME} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

- MAE (Mean Absolute Error):

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

- MPE (Mean Percentage Error):

  $$\text{MPE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right) \times 100$$

- RMSE (Root Mean Squared Error):

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

- ACF1 (Autocorrelation of First Order):

  $$\text{ACF1} = \frac{\text{Cov}(y_t, y_{t-1})}{\sqrt{\text{Var}(y_t) \cdot \text{Var}(y_{t-1})}}$$

  where: y_t represents the observed value at time t, y_{t-1} represents the observed value at time t-1, $Cov(\cdot)$ represents the covariance function, $Var(\cdot)$ represents the variance function.

- Correlation:

  $$\text{Corr} = \frac{\text{Cov}(y, \hat{y})}{\sqrt{\text{Var}(y) \cdot \text{Var}(\hat{y})}}$$

  where: y represents the observed values, $\hat{y}$ represents the predicted values, Cov(·) represents the covariance function, Var(·) represents the variance function [22].

- Min-Max Error:
$$\text{Min-Max err} = \frac{\max(|y_i - \hat{y}_i|)}{\max(y) - \min(y)}$$

A.8 Other Machine Learning Models for Time Series Forecast

- Long Short-Term Memory (LSTM)

  LSTM is a type of recurrent neural network (RNN) that is designed to capture long-term dependencies in sequential data or time series. LSTM networks have a unique architecture that allows to selectively retain or forget information over different time steps [33]. This capability makes them well-suited for modeling and predicting complex temporal patterns [34].

- Prophet Model

  The Prophet model is a forecasting model developed by Facebook's Core Data Science team. It is specifically designed for time series forecasting with an emphasis on capturing seasonal patterns, trends, and holiday effects. Prophet incorporates several components, including piecewise linear trends, seasonalities, and custom seasonal effects. It also handles missing data and outliers using interpolation methods [33].

- Gradient Boosting Regression Trees (GBRT)

  GBRT is a machine learning technique that combines multiple decision trees to create a strong predictive model. It works by iteratively building an ensemble of decision trees, where each subsequent tree corrects the mistakes of the previous trees. GBRT can capture complex nonlinear relationships and interactions among features [27].

- SARIMAX

  SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables) extends the traditional ARIMA model by incorporating seasonal components and external variables [28] [29] [33]. SARIMAX models can capture the autocorrelation, trend, and seasonality in time series data, making them effective for forecasting. They use past observations and their lags to predict future values while considering the impact of seasonal patterns and selected variables [34] [33].

A.9 Style and Notation

- Language style: The document has been written in British English. As a result, certain spelling and terminology may differ from other English versions.
- Footnotes: The purpose of footnotes in this study is to provide further information and references to observatory data or data sources.
- References: The references in this study are mostly theoretical or documented articles published in academic journals that help with understanding of the topic.

A.10 Python code notebooks

The code notebooks used can be found in the following Git Hub Repository:

https://github.com/bernatm/ABDataChallengeExtension

References

[1] Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust Statistics: Theory and Methods.* Wiley.

[2] Ali S. Hadi & Jeffrey S. Simonoff. (1993). Procedures for the Identification of Multiple Outliers in Linear Models. *Journal of the American Statistical Association, 88*(424), 1264-1272. https://doi.org/10.1080/01621459.1993.10476407

[3] Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences.*

[4] Zainuddin, Aznilinda & Hairuddin, Muhammad & Yassin, Ahmad & Abd Latiff, Zatul Iffah & Azhar, Aziemah. (2022). *Time Series Data and Recent Imputation Techniques for Missing Data: A Review.* 346-350. https://doi.org/10.1109/GECOST55694.2022.10010499

[5] Xi Wang, & Chen Wang. (2004). *Time Series Data Cleaning with Regular and Irregular Time Intervals.* School of Software, Tsinghua University.

[6] Imani, S. (2021). *Multi-Window-Finder: Domain Agnostic Window Size for Time Series Data.*

[7] Hassani, H., Kalantari, M., & Ghodsi, Z. (2019). *Evaluating the Performance of Multiple Imputation Methods for Handling Missing Values in Time Series Data: A Study Focused on East Africa, Soil-Carbonate-Stable Isotope Data.* Stats, 2(4), 457–467. https://doi.org/10.3390/stats2040032

[8] Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data.* Prentice-Hall, Inc.

[9] Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). *An extensive comparative study of cluster validity indices.* Pattern recognition, 46(1), 243-256.

[10] Bishop, C.M. (2006) *Pattern Recognition and Machine Learning.* Springer, Berlin.

[11] Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysi*s. John Wiley & Sons.

[12] Rousseeuw, P. J. (1987). *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.* Journal of Computational and Applied Mathematics, 20, 53-65.

[13] Calinski, T., & Harabasz, J. (1974). *A dendrite method for cluster analysis.* Communications in Statistics, 3(1), 1-27.

[14] Davies, D. L., & Bouldin, D. W. (1979). *A cluster separation measure.* IEEE Transactions on Pattern Analysis and Machine Intelligence, (2), 224-227.

[15] Royston, P. (1983). *A remark on algorithm AS 181: The W-test for normality.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 32(2), 212-215.

[16] Razali, N. M., & Wah, Y. B. (2011). *Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests.* Journal of Statistical Modeling and Analytics, 2(1), 21-33.

[17] Lehmann, E. L., & D'Abrera, H. J. M. (2018). *Nonparametrics: Statistical methods based on ranks.* Springer.

[18] Stevens, S. S. (1946). *On the theory of scales of measurement.* Science, 103, 677–680. https://doi.org/10.1126/science.103.2684.677

[19] Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). Wiley-Interscience.

[20] Mahbod, M. H. B., Chng, C. B., Lee, P. S., & Chui, C. K. (2022). *Energy saving evaluation of an energy efficient data center using a model-free reinforcement learning approach.* Applied Energy, 322, 119392.

[21] Willmott, C. J., & Matsuura, K. (2005). *Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance.* Climate research, 30(1), 79-82. doi: 10.3354/cr030079

[22] Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice.* OTexts.

[23] Hartwig, F. P., Davey Smith, G., Schmidt, A. F., Sterne, J. A. C., Higgins, J. P. T., & Bowden, J. (2020). *The median and the mode as robust meta-analysis estimators in the presence of small-study effects and outliers.* Research synthesis methods, 11(3), 397–412. https://doi.org/10.1002/jrsm.1402

[24] Huber, P. J. (1996). *Robust statistical procedures.* Society for Industrial and Applied Mathematics.

[25] Rousseeuw, P. J., Hampel, F. R., Ronchetti, E. M., & Stahel, W. A. (2011). *Robust statistics: the approach based on influence functions.* John Wiley & Sons.

[26] Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control.* John Wiley & Sons.

[27] Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine.* Annals of Statistics, 29(5), 1189-1232.

[28] Usha, T. and Balamurugan, S. (2016). *Seasonal Based Electricity Demand Forecasting Using Time Series Analysis.* Circuits and Systems, 7, 3320-3328. doi: 10.4236/cs.2016.710283.

[29] T. Jakaša, I. Andročec and P. Sprčić (2011). *Electricity price forecasting — ARIMA model approach.* 8th International Conference on the European Energy Market (EEM), Zagreb, Croatia, 222-225

[30] Lundberg, S.M., and Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions.* In Advances in Neural Information Processing Systems 30 (NIPS 2017), eds. I. Guyon et al., 4765-4774.

[31] Lundberg, S.M., et al. (2020). *From Local Explanations to Global Understanding with Explainable AI for Trees.* Nature Machine Intelligence, 2(1), 252-263.

[32] Schaffer, A.L., Dobbins, T.A. & Pearson, SA. (2021). *Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: a guide for evaluating large-scale health interventions.* BMC Med Res Methodol 21, 58.

[33] Ning, Y., Kazemi H., Tahmasebi P. (2022). *A comparative machine learning study for time series oil production forecasting: ARIMA, LSTM, and Prophet.* Computers & Geosciences,

[34] Zhou, Y. (2021). *Regional energy consumption prediction based on SARIMAX-LSTM model.* Academic Journal of Computing & Information Science, 4(3), 41-51.