Extracting temporal patterns from smart city data

Regina Gubareva - a49211

Thesis presented to the School of Technology and Management in the scope of the Master in Informatics.

Supervisors:

Prof. Rui Pedro Lopes

Prof. G. S. Burnakulova

This document does not include the suggestions made by the board.

Bragança

2021-2022

Extracting temporal patterns from smart city data

Regina Gubareva - a49211

Thesis presented to the School of Technology and Management in the scope of the Master in Informatics.

Supervisors:

Prof. Rui Pedro Lopes

Prof. G. S. Burnakulova

This document does not include the suggestions made by the board.

Bragança

2021-2022

Dedication

This thesis is dedicated to my supervisor and mentor, Professor Rui Pedro Lopes, who guided me through all of the challenges associated with my graduation. I will always be grateful for all of the efforts made to help me develop my technical skills, scientific thinking, and curiosity, and I will never loose my enthusiasm and perseverance.

In addition, I would like to thank the professors at my alma mater, Tomsk State University of Control Systems and Radio-electronics for instilling in me a passion for science.

To my home university, Taraz State University and the instructors who assisted and encouraged me.

Acknowledgment

This work is developed within the project “PandIA - Management of Pandemic Social Isolation Based on City and Social Intelligence”, supported by the FCT within the project scope DSAIPA/AI/0088/2020, which focuses on providing detailed information, such as resource consumption trends, estimation of people in each area or household, a heatmap of suspected disease outbreaks, and, overall, to assist health, municipal and emergency authorities in decision making.

Abstract

In the modern world data and information become a powerful instrument of management, business, safety, medicine and others. The most fashionable sciences are the sciences which allow us to extract valuable knowledge from big volumes of information. Novel data processing techniques remains a trend for the last five years, in a way that continues to provide interesting results. This paper investigates the algorithms and approaches for processing smart city data, in particular, water consumption data for the city of Bragança, Portugal. Data from the last seven years was processed according to a rigorous methodology, that includes five stages: cleaning, preparation, exploratory analysis, identification of patterns and critical interpretation of the results. After understanding the data and choosing the best algorithms, a web-based data visualizing tools was developed, providing dashboards to geospatial data representation, useful in the decision making of municipalities.

Keywords: Data analysis, clustering, big data, water consumption

Аннотация

В современном мире данные и информация стали одним из самых мощных инструментов в управлении, бизнесе, безопасности, медицине, науке и социальной сфере. Самыми модными и востребованными науками в настоящий момент являются науки, позволяющие извлекать полезные знания из больших объемов информации. Новые методы обработки данных остаются тенденцией последних пяти лет и продолжают генерировать интересные результаты. В данной работе исследуются алгоритмы и подходы для обработки данных "умного города", в частности, данных о потреблении воды в городе Браганса, Португалия. Данные за последние семь лет обрабатывались в соответствии со строгой методологией, включающей пять этапов: очистка, подготовка, исследовательский анализ, выявление закономерностей и критическая интерпретация результатов. Цель исследованиия - определение шаблонов поведения в потрблении воды связанных с определенными событиями и построение модели прогнозова на основе найденных закономерностей. В результате исчерпывающего анализа с помощью множества методов было установлено отсутствие систематических зависимостей в рассматриваемом типе данных. На заключительном этапе был создан инструмент визуализации данных, обеспечивающий динамические панели для представления аналитических данных о распределении потребления. Разработанный инструмент управления аналитикой полезен для принятия решений муниципалитетом.

Ключевые слова: Анализ данных, кластеризация, большие данные, водопотребление

Contents

List of Figures

Acronyms

ARIMA Autoregressive Integrated Moving Average. 7, 10, 15, 16

BD-CVI Big Data Clustering Validity Indices. 11

CUSUM CUmulative SUM. 9, 15, 16

DBSCAN Density-Based Spatial Clustering of Applications with Noise. xii, 21, 54, 55, 57, 58, 59, 60, 61, 62

DNN Deep Neural Network. 7, 15, 16

EBC Edge-Betweenness Centrality. 11

ES Exponential Smoothing. 10, 11

FCM Fuzzy c-Mean. 12, 13, 16, 21

FIESTA-IoT Federated Interoperable Semantic IoT/Cloud Testbeds and Applications. 6

GK Gustafson–Kessel. 13, 21

GN Girvan–Newman. 11, 13, 17

GRNN General Regression Neural Network. 10, 15

HEBC Hyperbolic Edge Betweenness Centrality. 11

HGN Hyperbolic Girvan–Newman. 17

HIV Human Immunodeficiency Virus. 6, 10

IoT Internet of Things. 1, 4, 6, 11, 17

LSTM Long-Short Term Memory. 7, 10, 15, 16, 17, 21

MAE Mean Absolute Error. 13

MAPE Mean Absolute Percentage Error. 13

MSE Mean Square Error. 10, 13

OGC Open Geospatial Consortium. 10

PCA Principal Component Analysis. xii, 55

RMSE Root-Mean-Square Error. 13

t-SNE T-distributed Stochastic Neighbor Embedding. xii, 55, 58, 59, 60, 61

VAE Variational AutoEncoder. xii, 31, 55, 57, 58, 62

VAUD Visual Analyzer for Urban Data. 15

WSN Wireless Sensor Networks. 5

Chapter 1

Introduction

The collection and storage of heterogeneous and multi-source data has become prevalent in most of the society organization. The ubiquity of Internet of Things (IoT) devices and sensor networks, urbanization, and digitalization have served as the reason for developing a new direction in Big Data within the Smart City. Smart City is the urban management concept resulting from the combination of information, communication technologies, and the IoT, which provides effective administration and a higher quality of life.

The assessment of the feasibility of applying data analysis to the understanding or prediction of specific events in a specific locality, the impact on the social environment, and the relationship between water consumption and regional economic growth is fundamental to support social and economic growth, as well as contributing to the health and well-being of the population.

Overall, the significant meaning of the efforts made is the extraction of knowledge about the distribution of resource expenditures. The comprehensive mining analytics will be presented in the following chapters, which can be helpful for further investigations as a stack of important conclusions and of confirmed and unconfirmed hypotheses.

1.1 Context

This work is developed within the project "PandIA - Management of Pandemic Social Isolation Based on City and Social Intelligence", supported by the FCT within the project scope DSAIPA/AI/0088/2020, which focuses on providing detailed information, such as resource consumption trends, estimation of people in each area or household, a heatmap of suspected disease outbreaks, and, overall, to assist health, municipal and emergency authorities in decision making. For that, it uses information from several sources, including pathogen characteristics, infection statistics, municipal information, social networks, and hospital information and statistics [1].

1.2 Objective

Among the multitude of information, water consumption data provides a promising starting point for data analysis, not only because it mirrors the behaviour of the society (in all areas), but also because it is a fundamental resource that has to be correctly and rationally managed. In this work, large data sets of water consumption were analyzed towards discovering inherent or hidden consumption patterns. So, the objective of this work is to understand the structure, features, patterns, and changes of the water consumption data, obtained from Bragança municipality, and figure out how decision making can be improved based on the analysis results.

The second contribution is the development of interactive visualizing tools to simplify the perception of the regularities. High-quality data visualization is critical for data analysis and data-driven decision making. Visualization allows to quickly and easily realize and interpret associations and identify evolving trends that would not attract attention as raw data. The instrument is provided as a web-application based on the Apache Superset framework. This allows people without special training to visualize and understand the data.

1.3 Organization of the document

This document is organised in six chapters, starting with this introduction. Chapter 2 provides an explanation of context, used technologies and considered datasets. Chapter 3 describes the problematic, the main approaches, the requirements to instruments and the difficulties connected to subject. Chapter 4 presents the analysis process exhaustively and provides results interpretation. Chapter 5 defines the web-application development process and chapter 6 ends with some conclusions and final considerations.

Chapter 2

Context and Technologies

The smart city concept is popular and common in the scientific literature, characterizing a healthy environment that improves the quality of life and well-being of citizens [2]. Due to the diversity of services, resources and projects, smart cities manage huge amounts of data, typically within the Big Data concept. One can argue what are the minimum conditions and characteristics for a city to become "smart". However, since nowadays most operations are controlled via comprehensive information and communication technologies, the need to collect, store, integrate, process and analyze data is prevalent and important in most cities.

Over the past ten years, the number of sensors and metering devices has been increasing geometrically. The intention to control and understand everything surrounding us became a significant step in the development of technologies of environmental sensors: smart houses, smart cities, smart devices, IoT, and many others. Legacy information is also laying around, in spreadsheets or databases, that can be valuable if correctly accessed and integrated. Citizens and institutions also make use of social networks to convey opinions, criticism or information about resources, services or events. The essential questions are how to use this data and how to extract practical and meaningful information from all these measurements?

Big data is a set of technologies for processing large amounts of data. It refers not only to the amount of information, but also to the "data rate", meaning the multiple

streams of data that should be processed in real-time. Moreover, large examples of data usually enclose hidden, potential valuable, patterns. Several unique phenomena associated with high dimensionality, including noise accumulation, spurious correlation, and random endogeneity, makes traditional statistical procedures difficult to use. In the Big Data era, large sample sizes allow us to better understand heterogeneity by shedding light on research such as examining the relationship between specific covariates and rare outcomes.

2.1 Big Data

The smart city concept implies integrating multiple information and communication technologies for city infrastructure management: transport, education, health, systems of housing and utilities, safety, etc. Municipal governments collect numerous heterogeneous information, and an "urban data" term can mean various datasets: data from video surveillance cameras, traffic, air quality, energy and water consumption, images for smart recognition. Therefore for this study, the essential is to recognize and classify different datasets utilized in considered resources.

Trilles et al. describe a methodology of (big) data process produced by sensors in real-time [3]. It assumes that it works with different sensor data sources with different format and connection interfaces. Wireless Sensor Networks (WSN) are used for monitoring the physical state of the environment: air pollution, forest fire, landslide, water quality. Although the system proposed by the authors is designed to process all data types, the WSN mainly produce numerical data like water level, the gas concentration in air, mainly classified as quantitative information.

An efficient method to derive spatio-temporal analysis of the data, using correlations was proposed by [4]. The authors use data from bluetooth sensors installed in light poles. The data collected from the road sensors in the city of Aarhus in Denmark. The measurements are taken every 5 minutes and dataset includes timestamp, location information, average speed and a total of automobiles at the time of commit. The data were classified like numerical as there are no text, images, sound or video information.

Bordogna et al. used in their paper big mobile social data, which are included users-generated, geo-referenced and timestamped contents [5]. The content means text data that users posts in the modern emerging social systems like Twitter, Facebook, Instagram, and so forth. Hereby, the dataset can be classified as heterogeneous by way of contains text of social networks posts and numerical data of location and time.

Wang et al. considered another approach to analysis and evaluated the effectiveness of deep neural networks [6]. The aim of their paper was the monitoring and control of local Human Immunodeficiency Virus (HIV) epidemics. The collection includes statistics on the number of morbidities, mortality and mortality by region, age, sex, occupation. The type of data categorized as text and numerical.

The researchers from Spain, Pérez-Chacón et al., proposed a methodology to extract electric energy consumption patterns in big data time series [7]. The study used the big data time series of electricity consumption of several Pablo de Olavide University buildings, extracted using smart meters over six years.

Karyotis et al. presented a novel data clustering framework for big sensory data produced by IoT applications [8]. The dataset was collected from an operational smart-city/building IoT infrastructure provided by the Federated Interoperable Semantic IoT/-Cloud Testbeds and Applications (FIESTA-IoT) testbed federation. The array is heterogeneous and represents measurements of different types: temperature, humidity, battery level, soil moisture, etc.

Azri et al. presented a technique of three-dimensional data analytics using a dendrogram clustering approach [9]. It is assumed that the algorithm can be applied to large heterogeneous datasets gathered from sensors, social media, and legacy data sources.

Alshami et al. tested the performance of two partition algorithms K-Means and Fuzzy c-Mean for clustering big urban datasets [10]. Compared techniques can be applicable for huge heterogeneous datasets in various areas like medicine, business, biology, etc. In the paper, the authors utilized the urban data from various data sources, such as Internet of Things, LIDAR data, local weather stations, mobile phone sensors.

Chang et al. developed a new iterative algorithm, called the K-sets+ algorithm for

clustering data points in a semi-metric space, where the distance measure does not necessarily satisfy the triangular inequality [11]. The algorithm is designed for clustering data points in semi-metric space. To understand what semi-metric space is, it is necessary to briefly consider the concept of metrics in space. The metric is the mapping for some set $d : X \times X \to \mathbb{R}$, for which the axioms of non-degeneracy and symmetry have to be satisfied but not necessarily the triangle inequality. If the distance between different points can be zero, the metric is semi-metric. The method was evaluated with two experiments: community detection of signed networks and clustering of real networks. The dataset included 216 servers in different locations, and the latency (measured by the round trip time) between any two servers of these 216 servers is recorded in real time.

Chae et al. have compared the performance of the Deep Neural Network (DNN), Long-Short Term Memory (LSTM), the Autoregressive Integrated Moving Average (ARIMA) in predicting three infectious diseases [12]. The study uses four kinds of data to predict infectious disease, including search query data, social media big data, temperature, and humidity. Data related to malaria, chickenpox, and scarlet fever, for 576 days, were considered. As a result, the data is partly numerical and partly is text.

The research of Chen et al. focuses on multi-source urban data analysis [13]. The points of interests are the geographical, street view, road map and real-estate data. The record comprises road network of the city, longitude, latitude, name, and functionality of a structure in the urban environment, imagery of locations. Obviously, the dataset is ranked as heterogeneous.

Simhachalam and Ganesan presented a multidimensional mining approach in a successive way by finding groups (clusters) of communities with the same multi-dynamic characteristics [14]. The data refers to the statistics of population, migration, tax capacity, dwellings, employment, and commuters.

The majority of the studies assume heterogeneous nature data. There are two research papers with only numerical data and one of the paper investigates image data processing. Text and numerical data are dominant and they are collected from multiple sources (Table 2.1).

Paper	Data	Category
[3]	data from different sensors	heterogeneous
[4]	traffic data collected from the road sensors in the city: geographical location, time-stamp, average speed and total of automobile	numerical
[5]	social networks posts, timestamp, geo-location	heterogeneous
[6]	10-year historical HIV incidence data: the number of morbidity, morbidity, mortality and mortality by region, age, sex, occupation	heterogeneous
[7]	electricity consumption for 6 years for several buildings	numerical
[8]	big sensory data, measurements of different types: temperature, humidity, battery level, soil moisture	heterogeneous
[9]	smart city data	heterogeneous
[10]	data from Internet of Things, LIDAR data, local weather stations, mobile phones sensors	heterogeneous
[11]	locations and the latency (measured by the round trip time) between any two data points	heterogeneous
[12]	search query data, social media big data, temperature, and humidity	heterogeneous
[13]	geographical data, points of interests data(longitude, latitude, name, and functionality of a structure in the urban environment), street view data, real estate data, mobile phone location data, social network data, micro-blog data, taxi GPS trajectory data, taxi profile data	heterogeneous
[14]	the measurements of the blood tests as corpuscular volume of test substances and the number of half-pint equivalents of alcoholic beverages drunk per day	numerical

Table 2.1: Data types and sources.

2.2 Algorithms

Following the literature introduced before, it matters to check the algorithms that are usually used in big data and smart city. In general, 12 different approaches to big municipal data processing were found. The methods can be divided into groups depending on the manner of information handling: clustering, classification, correlation, deep neural network, frameworks, and community detection (Figure 2.1).

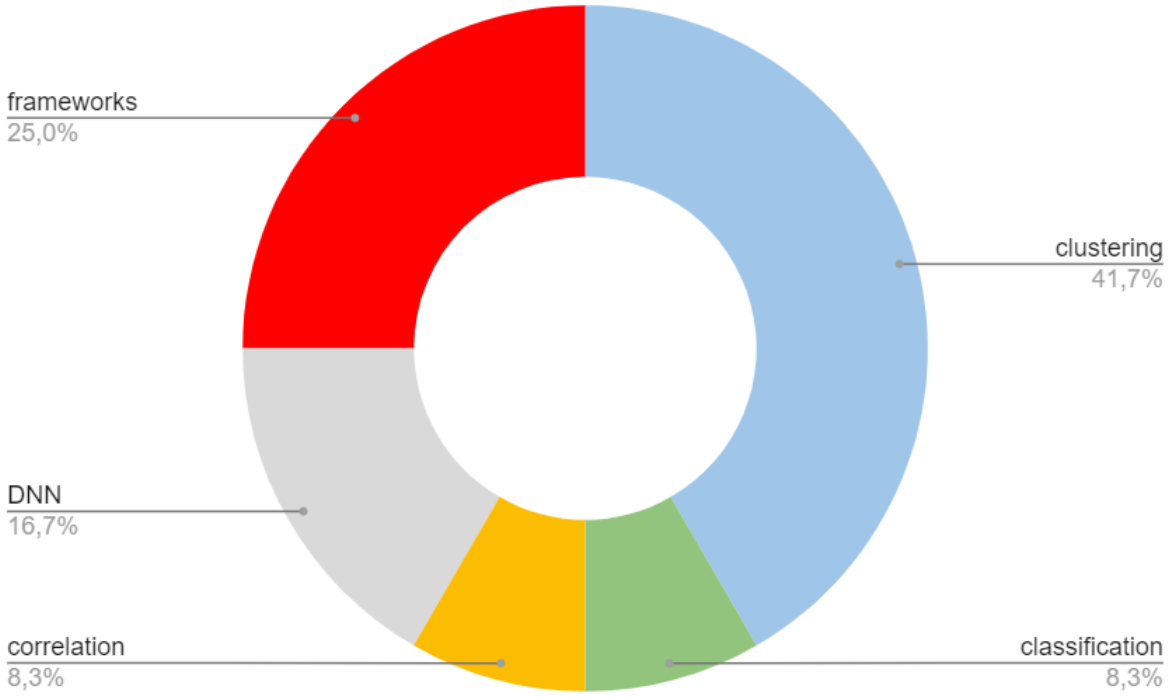

Figure 2.1: Methods of urban data analysis

The approach followed by [3] includes three layers: content layer, services layer, and application layer. The content layer includes sensor networks data sources, and the services layer provides database connection, transformations of data, communications protocols for real-time data handling and processing. The last layer implies client application. The service layer implements CUmulative SUM (CUSUM) algorithm of anomalies detection. The method considers the set of observations following normal distribution. For each collection of measurements, the cumulative sum is calculated. When the score overcomes

the threshold, the algorithm detects anomalies. If the parameter exceeds the threshold, the anomaly will be due to the increase (up-event), and if the sum is greater than the threshold, it will be due to the decrease (down-event). Different data types from multiple sources are processed by a special wrapper and transformed to standard form. Transformed observation is encoded in line according to Open Geospatial Consortium (OGC) standard for Observations and Measurements.

The unique method used in [4] tried to apply correlations methods to urban data analysis. They suggested an efficient method to derive spatio-temporal analysis of the data, using correlations, with Pearson and Entropy-based methods and compares the results of both algorithms. Pearson's coefficient characterizes the presence of linear dependence between two values. The weakness of Pearson correlation is poor accuracy when variables are not distributed normally. Mutual information is the statistical function of two random variables, which describes the quantity of information of one random value in another. The constraint of mutual information is that it has a higher processing complexity than Pearson correlation. The technique continuously calculates the average correlation for sensory road data divided into two sectors until the data runs out. Different types of correlation were tested.

[6] used LSTM neural network models, ARIMA models, General Regression Neural Network (GRNN) models and Exponential Smoothing (ES) models to estimate HIV incidence in Guangxi, China, and explore which model is the best and most precise for local HIV incidence prediction. ARIMA is the model used for time series forecasting. LSTM is a recurrent neural network, characterized by the ability to learn long-term dependencies. In this study, several models were built. The model with the lowest Mean Square Error (MSE) was considered the optimal model. GRNN is a feedforward neural network, which estimates values for continuous dependent variables. The principal advantages of GRNN are fast learning and convergence to the optimal regression surface as the number of samples becomes very large. GRNN is particularly advantageous with sparse data in a real-time environment because the regression surface is instantly defined everywhere, even with just one sample. The method is usually used for functions' approximation, so

it can provide very high accuracy, but for huge samples is computationally expensive. ES model is one of the simplest and widespread practices of series alignment. The method can be presented as a filter that receives the original series members as the input, and the output forms the current values of the exponential average.

[7] search patterns in data related to electricity consumption. The methodology describes all stages of data processing: data collection, cleaning, transformation, index analysis, clustering, and results. The first stage aims to pre-process the data so that they can be clustered. The second phase consists of obtaining the optimal number of clusters for the dataset by analysing and interpreting various cluster validation indices. Next, K-Means is used for clustering and, finally, retrieve the centroids for each cluster. The processing is done in Apache Spark and the algorithms include Big Data Clustering Validity Indices (BD-CVI) and K-Means.

[8] modified the community detection algorithm Girvan–Newman (GN) [15] algorithm for big data clustering of IoT sensors. Their method organises complex data in blocks, called communities or modules, according to certain roles and functions, organized in a multigraph. The problem is to find in a given multigraph a partition of vertices where the objective function is minimized. To achieve this, the graph edges are deleted iteratively, depending on the value of the metric. The Edge-Betweenness Centrality (EBC) is the most common metric used, but the computation for this is time-consuming. The authors suggested a new measure approximating EBC, which capitalizes on hyperbolic network embedding and can be considered as the "hyperbolic" analog of EBC. This measure is denoted as Hyperbolic Edge Betweenness Centrality (HEBC), and it is computed by utilizing the hyperbolic node coordinates assigned to the embedded nodes. The novel metric enhances the performance without harming accuracy.

The other technique of data organizing and processing proposed by [9] and implies 3D data analytics using dendrogram (hierarchical) clustering approach. 3D data represents a structure of information that combines, simultaneously, the classification and clustering tasks. The organized data is mapped to tree structure and retrieved by tree traversal algorithms. Dendrogram clustering is a method of merging objects into bunches. In the

study, the bottom-up algorithm of clustering is utilized, which means that each item in a class is assigned to a single cluster. Then combine the clusters until all objects are merged together. An important parameter is the distance between objects in a class. The metric shows a quantitative assessment of the items' similarity ratio according to different criteria. The given research does not provide a selection of the specific parameter, although the choice of metric occurs on the second step of the method. The ability to retrieve information and the efficiency of the structure were measured. In general, the technique demonstrates a good characteristic of information extraction but not the most attractive performance parameters.

Other clustering algorithms, Fuzzy c-Mean (FCM) and K-Means, were tested by [10]. K-Means algorithm is one of the simplest methods but at the same time the most inaccurate. The main idea is that at each iteration, the center of mass is recalculated for each cluster obtained in the previous step, then results are partitioned into clusters again under new centers. The algorithm ends when the cluster is not changed in iteration. The fuzzy c-Mean method allows for obtaining "fuzzy" clustering of large sets of numerical data and makes it possible to correctly identify objects at the boundaries of clusters. However, the execution of this algorithm requires serious computational resources and the initial setting of the number of clusters. In addition, ambiguity may arise with objects remote from the centers of all clusters.

[11] designed a new approach for clustering data points. In essence, the method is an extension of the K-set clustering algorithm for semi-metric space. The problem of the K-sets approach is that triangle distance is not non-negative. Thus the K-sets algorithm may not converge at all and there is no guarantee that the output of the K-sets algorithm are clusters. For solving this difficulty, the definition of triangle distance was adjusted, so that the non negativity requirement could be lifted. The experimental results confirm the proficiency of the method for the geographic distance matrix and the latency matrix.

[12] used deep neural networks for the prediction of infectious diseases.

[13] presented a visual analysis framework for exploring and understanding heterogeneous urban data. A visually assisted query model is introduced as a foundation for

interactive exploration coupled with simple, yet powerful, structural abstractions and reasoning functionalities.

One more clustering method is used by [14]. FCM, K-Means, and Gustafson–Kessel (GK) clustering algorithms are implemented. According to the paper the most accurate and effective algorithm is K-Means clustering, but the other methods have its own advantages and show higher correctness in certain cases.

2.3 Complexity Assessment

The algorithms described above have characteristics of performance and scalability that should be understand. Table 2.2 gives a comprehensive description of the complexity and accuracy of the considered algorithms. For many, it was not easy to evaluate the complexity since the time depends on the characteristics of the machine. Therefore rough estimates are provided, and all presented assessments are for worst case values.

The K-sets+ algorithm yields the highest performance from all cluster algorithms. The time complexity is linear $O((Kn + m)I)$, where I is the number of iterations. The other method with linear time is FCM with $O(nCI)$, where C - number of clusters, I - number of iterations. If we compare exponent for these two approaches, the apparent fact is that the K-sets+ gives a little advantage. Considering the accuracy, K-sets+ has 95% as the worst result. The FCM algorithm gives the complexity on average 81,97%. It is noteworthy that the GN modification provides 100% accuracy for most datasets and only 50% in the case of outliers. It could be used for a dataset with low sparseness if high accuracy is required. The dendrogram clustering method is slower than the others but can produce a hierarchical tree structure for data. K-Means clustering is the simplest method but has a quadratic complexity and an accuracy of not more than 88% for different input data.

The deep learning algorithms were compared by the set of parameters: MSE, Root-Mean-Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). From the results given by the authors, it follows that the most accurate

Algorithm	**Purpose**	**Complexity**	**Accuracy**	**Papers**
CUSUM algorithm	anomaly detection	O(n)	—	[3]
Mutual information and Pearson correlation	find correlation between sensory data	$O(n^3)$	mutual information shows higher accuracy	[4]
LSTM	predict diseases	the computational complexity per time step is $O(w)$, where w is a weight for one cell	more than 85%	[6]
ARIMA	predict diseases	—	80%	[6]
GRNN	predict diseases	—	76%	[6]
ES	predict diseases	—	74%	[6]
K-Means clustering	to extract electric energy consumption patterns	$O(n^2)$	78%	[7]
modification of Girvan-Newman algorithm	community detection	$O(n^2)$	on average 65%, the highest 100% when $k = 3$	[8]
Dendrogram clustering	produce hierarchical tree structure for data for data retrieval and analytics	$O(n^3)$	—	[9]
K-Means	clustering	$O(n^2)$	87,94%	[7], [10]
Fuzzy c-Mean	clustering	$O(nCI)$, C - number of clusters, I - count of iterations	81,91%	[10]
K-sets+	clustering in metric space	$O(Kn + m)$	95%	[11]
DNN	predicting infectious diseases	n examples learning complexity, repeated k times, is $O(wnk)$, where w is number of weights	77%	[12]
VAUD	spatio-temporal data visualisation	—	78,6% in worst case	[13]
K-Means	clustering	70,22%	87,94% the best, 60,41% the lowest	[14]
Fuzzy c-Means (FCM)	clustering	68,54%	81,91% the best, 56,25% the lowest	[14]
Gustafson–Kessel (GK)	clustering	60,68%	95,83% the best, 66,19% the lowest	[14]
Similarity-Matrix-based Clustering	trip clustering	$O(n^3)$ - the worst case	—	[5]

Table 2.2: Algorithms assessment

algorithm is LSTM, but at the same time, the slowest. The fastest method is ES, but with the worst accuracy. All deep learning algorithms were used for predicting diseases. The accuracy of the ES and GRNN model was relatively poor [6]. The ARIMA model has several requirements: the time series should be stationary with steadily changing differences, and only linear relationships could be captured [6]. The DNN and LSTM models were observed to be sensitive to decreasing trends and increasing trends, respectively [12]. It worth to notice that the time complexity for deep neural network is hard to evaluate with O notation. The authors provide real time results, according with the considered researches the fastest model is ES, and the slowest is LSTM.

The CUSUM algorithm is time linear complexity. The solution is straightforward and fast but has limitations that must be taken into account, such as the consideration that all the series must follow a normal distribution and a series of observations cannot have trends [3]. Visual Analyzer for Urban Data (VAUD) presents the visualising of heterogeneous urban data. The approach is based on queries to the database, hence the time complexity can not be estimated. The data gathered from mobile phones and stored in one database combines different queries and different results are obtained. The accuracy on average for queries is 76%.

One of the widespread statistical methods applied to big data is a correlation. In the listed papers, there is one algorithm that considered the correlation applied to smart city data. The study compared two types of methods: Pearson correlation and Mutual information. The time complexity for both is a cube. But Pearson correlation can discover linear distribution of data, and mutual information can discover dependencies in more general data distribution cases. However, if an application prioritises real-time response over the accuracy, Pearson correlation will be suitable as it will only give a few false negatives. In other scenarios with different types of data streams (temperature, pollution, etc.), it is better to use mutual information without a priori knowledge of the potential correlations because we do not know the percentage of cases where Pearson correlation will fail to detect the correlations [4].

The assessment of time and accuracy of all proposed algorithms demonstrate that if

our purpose is prediction, the best variant for us is deep neural networks like LSTM. For effective clustering, the K-sets+ or FCM algorithms are the most powerful. If it is necessary to obtain additional analysis, it is possible to find the correlation. Considering the context of municipal data the frameworks are beneficial, as they assume all stages of data processing from storage to visualizing.

2.4 Outcomes and Purpose

This section provides a brief analysis of the outcomes and purpose of each paper.

In [3], the detected anomalies by CUSUM algorithms create the warning message for the client-side in the case of rare events. Each event contains a sensor identifier (sender field) and the identifier of the particular observation that has caused the event (identifier field). An event dashboard visualizes this data. The panel shows all sensing nodes of a network on a map using markers. Inside each marker, the amount of events that have been detected for this particular sensing node appears. If this node triggers an event, the marker turns red, than not the marker remains blue.

The analyses based on the correlation and mutual information were used to monitor the traffic of the city. Three sets of experiments have been performed. In the first one, the performance of Pearson correlation and mutual information was compared [4]. The results were visualized on Google Maps. It can be concluded that the Pearson correlation is effective for linear distribution of data, and mutual information is vital for nonlinear dependencies but requires more time.

The results obtained by [6] are predictions of HIV disease for two years. Each compared algorithm has its metrics. For example, ARIMA includes a moving average process, an autoregressive moving average process, an autoregressive moving average process, and an ARIMA process according to the different parts of the regression and whether the original data are stable. To evaluate data accuracy, they compared with original information about HIV cases for 2015 and 2016 years. The same type of outcomes data demonstrates the [12]. They compared the same parameters for LSTM, DNN, and ARIMA to evaluate

infectious disease prediction correctness.

All clusters algorithms give the same result as a count of clusters and their accuracy.

In [7], the electricity consumption data were clustered in 4 and then in 8 groups. The outcomes presented as diagrams. The clusters are categorized depending on buildings, seasons of the year and days of the week.

[8] provided the modification of the GN method with a novel metric. The proposed method was applied to multidimensional data obtained from an operational smart-city/building IoT infrastructure. The authors presented an accuracy evaluation, modularity and time comparison of Hyperbolic Girvan–Newman (HGN) and GN, comparing execution time of GN and HGN algorithms for graphs with known communities and modularity comparison for 5, 10, 20, 30 and 60-minute sampling. Given that statistics demonstrate the computational efficiency and that algorithm can give accurate outcomes.

[9] visualized the clusters into dendrograms, as tested on the information about 1000000 buildings. Response time analysis was provided as well, which exhibits that response time for the proposed method 50-60% faster than non-constellated data.

Data visualization framework presented by [13]

2.5 Summary

This chapter described the main trends in the city's infrastructure data processing, through a detailed analysis of the twelve papers. The authors considered the techniques of urban data processing. The input and output data, assessments of algorithms' effectiveness, and methods description are provided. The primary focus of the findings is deep neural networks in a way as the first intention of the research was creating or predicting models. From this point the most effective approach is leverage of LSTM. Based on surveyed articles LSTM gives the highest accuracy of prediction and is the fastest solution in comparison with similar solutions.

Chapter 3

Algorithms and Approach

The problem in the thesis can be clarified by three main points, namely, to make a comprehensive analysis of the big data massive containing the water consumption statistics for seven years, to produce the comparable analysis in terms of social and geographical factors, to understand the dependencies and assess the likelihood of predicting certain phenomena using expenditures analysis, and to create a web-application consisting of systematic analytical dashboards for decision-making.

Within these, there are also some transversal challenges to approach, such as analytical issues, appearance, usability, and functionality.

3.1 Challenges

The analytical issues are related to data collection, processing and storage, analysis, and result interpretation. All of these procedures are not linear and may result in logical, mathematical, or processing errors. The mathematical errors may be caused by statistical inaccuracy in handling methods such as “noise accumulation”, “false correlation” and “random endogeneity” [16]. Processing errors include accidental data loss, such as during cleaning or when a sample is insufficient for the given problem. Logical flaws typically assume one of two scenarios: the conclusion is based on an incorrect statement or the statement is correct, but the conclusion is incorrect. Given the massive amount of data

to analyze, some erroneous associations will emerge because of bias or other uncontrolled factors. If errors are rare and close to random, the final analysis's conclusions may be unaffected.

The appearance issue is linked to the proper visualization of outcomes. The value of visualization is determined by its ability to easily, efficiently, accurately, and correctly tell the story contained in the information. The charts and diagrams should be clear, visible, and accurately reflect the discussed situation. Color is important in visualizing information, particularly in interpersonal communication. Many people have peculiarities, if not disorders, in their perception of color and tone, which should be considered when selecting a color palette. This is especially true when creating final visual images for decision makers. The correct representation of information requires specialized knowledge and skills, as well as strong reliance on the developer's experience. The final result of visualization is frequently subjective for a specific person, but the main criteria of comfortable intelligibility must be met.

The ease of use of the instrument is reflected in its usability. Functionality issues are associated with a variety of features that must be implemented. When it comes to the application, functional issues usually arise when it does not work as intended initially. When it comes to data analysis and current work, these two parameters are related to the dashboard web-application.

The analytical web-application for decision-making is a single-page web-site that contains the collection of charts that provide comprehensive analytics for the major data frame. "Decision making" means that the person can draw conclusions about water consumption, calculate future needs, identify abnormal behavior, and be aware of the overall situation by analyzing the board. All basic analytics graphs, exploratory charts, and characterizations for all parameters should be included in the dashboard.

The initial goal of this investigation was to identify sequence patterns and identify the events that affected water usage in a given region, allowing for forecasting and predicting conditions based on resource expenditures. The data evaluation was conducted using the methodology depicted in Figure 3.1.

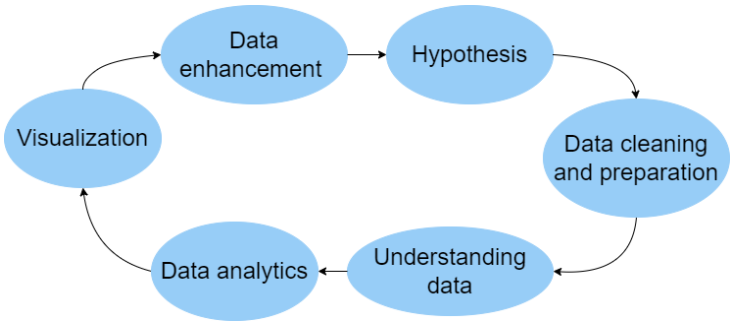

Figure 3.1: The data analysis cycle

The first step is to set a hypothesis or objective, determining the basic questions: what are we looking for? what are we supposed to discover? what is the goal of the analysis? The next step is to prepare and clean. At this stage, unnecessary records are eliminated for the established hypothesis and create a data frame that is most convenient for the task solving.

In the phase of understanding data, the preliminary analysis is performed to figure out the structure of data and recognize the methods for further examination. Data analytics assumes more detailed analysis that directly solves the specified task and either refutes or confirms the hypothesis.

Visualization is an obvious and essential step, as it makes it quick and easy to notice and interpret connections and regularities as well as identify developing trends that would not attract attention as raw data.

Further inquiries can be made after the results are interpreted and adjudicated. Primarily, the findings might meet the questions posed in the first iteration. Normally, it is a rare situation and it demands additional analysis. The results could indicate the presence of errors in the data frame and the need for additional information. All of these procedures are performed during the data enhancement phase.

The extraction of spatio-temporal patterns applied to resource consumption information is not a traditional approach. Water consumption data is typically considered in the context of urban data and smart cities and it is grouped together with expenditures for electricity, gas, and heat. Therefore, the related works are based on a review of the

approaches and trends in urban resource analysis. We looked at works that contained descriptions of various algorithms used in the subject. The central questions are: what methods are used for resource data analysis, what types of data are considered, and how effective of each technique?

The input and output data vary depending on the method and purposes of the research. Predominantly, heterogeneous data sources are considered. Images are used in one case, and numerical information is used in other two. Heterogeneous data refers to information derived from different sources, such as images, text, and numerical data. The reviewed papers make no use of video or audio data. The majority of the investigation provides a clustering model in terms of output results. Frameworks are the second most common. Deep neural networks, classification, and correlation models are among the remaining outcomes [1].

The findings are primarily focused on deep neural networks, and the original goal was to create or predict models. The most effective approach at this point is to leverage LSTM. According to the surveyed articles, LSTM has the highest prediction accuracy and is the fastest solution when compared to similar solutions. The learning rate of LSTM depends on the number of cells. As the following step of the study, it is planned to develop a predicting model based on city resources consumption data.

One of the key findings of the related work is that clustering is the most commonly used approach for numerical data. The most effective prediction method is LSTM. K-Means, FCM, K-sets, GK, Similarity-Matrix-based, and Dendrogram are some good clustering techniques with high efficiency. The author acknowledged the capability of clustering algorithms, but for the thesis, DBSCAN method was selected as the best for unstructured data. The next step was to use deep learning to forecast social phenomena.

The thesis presents a comprehensive analysis of water consumption as well as the development of a visualisation application. The investigation yielded a large number of results, each of which is interpreted.

3.2 Understanding the Data

One of the main tasks of this work is to research and extract valuable information from water consumption data. However, additional information is necessary to explain or check certain results of resource utilization. In general, there are five main datasets: yearly precipitation level data, population growth statistics, events and holiday records, supportive repositories, and COVID-19 infection statistics and policies. Each one is thoroughly examined, and the reason and necessity for their use are explained.

The data for the current investigation was provided by the municipality of Bragança, Portugal. The given statistic is real and represents actual water consumption by the city's population.

The main dataset is in tabular format and includes the following columns: year, month, installation number, consumer number, consumer type, installation zone, counter number, counter manufacturer, counter caliber, measure method, and consumption.

The data covers seven years: 2013, 2014, 2016, 2017, 2018, 2019, and 2020. Therefore, all records match these determined years and the study considers the period from 2013 to 2020 except 2015. The month value indicates the time when the consumption value was registered. Water and other resource usage is typically measured monthly, so using the year and month pair to identify periods of leverage resources is appropriate.

An installation number is the unique code for each counter device, corresponding to the address of the provision of the service. The counters can be installed in houses, apartments, factories, shops, industries, churches, and other urban facilities that are provided with city water. This feature is required for non-ambiguous recognition of all counter's installations, as there are numerous cases that can lead to controversial conclusions if the above parameter is not present: one consumer may have many installations in different locations, one object may have many counters, many objects may share the same counter. To identify unambiguously the counter and the object, the installation number is used.

A consumer number is the identification code assigned to a person who concludes an agreement with a water supplier. Every customer has a unique consumer number.

Nevertheless, it is important to highlight the fact that a single consumer can have many counters and installations for the reason that it will have an important influence on the whole analysis.

A consumer type is the class of object provided by water. There are 22 different types of consumers depending on the type of household that uses water (Table 3.1). As an example, type "DOMÉSTICO" means the private domestic household sector, where people use water for personal needs. There are industrial, rural, public utilization, and other types.

Table 3.1: Types of consumers

Consumer type	**Name**
1	DOMÉSTICO
2	COM/INDUSTRIAL/OBRAS
3	UTIL.PUBLICA
4	OBRAS
5	ESTADO
6	IGREJAS
7	EXP.A.RURAL
9	RURAL DOMÉSTICO
10	RURAL/ESTADO
11	FAM.NUMEROSAS
12	FAM.CARENCIADAS
13	NUMER./CARENC.
14	CP.DOM/URB
15	CP.COM/URB
16	CP.DOM/RURAL
17	CP.COM/RURAL
18	IPSS/IGR/RURAL
19	DOM./RURAL A.S
20	COM./RURAL A.S
21	REGA
22	CMB

The installation zone is connected with Bragança municipality geographical zones (Figure 3.2). The Bragança district includes 66 parishes (Table 3.2).

Portugal has 18 districts. one of them is Bragança. Each district consists of parishes, which is the smallest administrative division in Portugal. Regarding geographical zones,

Table 3.2: Geographical zones and its codes, Bragança, Portugal

Zona	**Nome**	**Zona**	**Nome**
1	Gimonde	43	Castrelos
3	Santa-Maria	44	Zoio
4	Samil	45	Faílde
6	Sé	46	Rabal
22	Crijó de Parada	47	S.Julião de Palacios
23	S.Pedro de Sarracenos	48	Deilão
24	Nogueira	49	Rio de Onor
25	Izeda	50	Alfaião
26	Pinela	51	Baçal
27	Castro de Avelãs	52	Santa Comba de Rossas
28	França	53	Gostei
29	Sortes	54	Paradinha Nova
30	Pombares	55	Serapicos
31	Rebordãos	56	Mós
32	Babe	57	Sendas
33	Quintela de Lampças	58	Milhão
34	Coelhoso	59	Donai
35	Outeiro	60	Macedo de Mato
36	Salsas	61	Quintanilha
37	Aveleda	62	Carrazedo
38	Rio Frio	63	Condesende
39	Parada	64	Rebordaínhos
40	Espinhosela	65	Calvelhe
41	Carragosa	66	Parâmio
42	Meixedo		

Figure 3.2: Geographical zones of Braganca region, Portugal

it can be seen that the central area, which includes Meixedo, Baçal, Sé, Santa Maria, and Gimonde is the territory of Bragança city. The geographical zone is represented in the dataset by a number from 1 to 66 and can be substituted by name according to Table 3.2.

The counter number is the unique identifier of the counter. The code is just a marker intended for determining the device and is a simple integer number.

The counter manufacturer is a string code that specifies the producer of the counter. There are various companies, this characteristic was not analyzed in detail for the reason of insignificance for further inquiries.

The counter caliber is directly related to the diameter of the installed device. The size of the meter indicates the number of water taps that can be handled. For example, a meter with a caliber of 20 millimeter powers 6 to 10 devices, while a 15 millimeter meter should power a maximum of 5 devices. The 15 millimeters gauge is the minimum to install, and given that it caters to the aforementioned maximum of 5 devices, it will fit a more modest home in terms of water usage. As a result, the counter calibre has no effect on consumption volume and is only meaningful from the position of indications recording, and thus has no influence on consumption patterns. This parameter can be neglected as a minor.

There are 8 different types of measuring consumption, indicated by big latin letters: A, D, H, L, M, N, T, V, depending on the type of the meter. Figure 3.3 illustrates the dependency between consumption and consumption reading type. The chart suggests that the majority of meters are of type L; consumption is increasing over time; counters of types V, N, and A are not used; and the number of counters of types T, M, L, and D remains stable. Simultaneously, it is absolutely obvious that the measurement method does not affect water expenditure.

The last column in the main dataset is the consumption. It normally contains an integer number denoting the monthly volume of water used in cubic meters.

As mentioned before, besides water consumption, additional information is also necessary, such as the yearly precipitation, Portugal's demography, and COVID-19 case datasets.

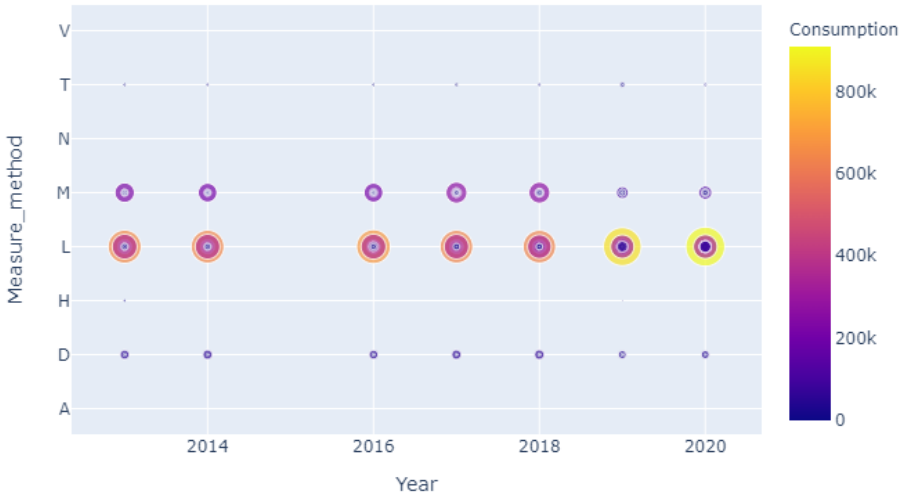

Figure 3.3: Water consumption per year for each metering method

Precipitation is water in the liquid and solid state that falls from clouds or settles from the air. The most common types of precipitation from clouds in Bragança are rain, drizzle, and, rarely, snow. There are several indicators to measure the amount of rainfall. The most common is expressed in millimeters (mm) and measured as one liter of water per square meter.

The records of precipitation are simple. They include 7 years considered in the paper, month, and level of rainfall in millimeters (Table 3.3).

Table 3.3: Sample of the precipitation level dataset

Year	**Month**	**QPRtot (mm)**
2013	1	5.504
2013	2	2.674
2013	3	7.023
2013	4	2.058
2013	5	0.915

In the current study, the research is not restricted by consumption patterns' extraction. Understanding the basis of changes in these patterns is equally important. The last year

in question fell during the pandemic of COVID-19. To provide a comprehensive picture, the disease case history was analyzed in relation to water expenditure. The dataset for COVID-19 cases encompasses only the Bragança region, 2020 year (Table 3.4).

Table 3.4: COVID cases in Braganca, 2020

Year	**Month**	**Cases**
2020	1	416
2020	2	423
2020	3	804
2020	4	2146
2020	5	3055
2020	6	2777
2020	7	1106
2020	8	887
2020	9	1008
2020	10	1658
2020	11	322
2020	12	794

The amount of water flowing depends directly on the population of the territory. Therefore, Bragança's population growth statistic has been taken into account. The dataset covers the population change rates for the last seven years, except 2015 (Table 3.5).

Table 3.5: Braganca's population changes

Years	**Net increase**	**Natural increase**	**Net migration**
2013	-302	-203	-99
2014	-340	-226	-114
2016	-267	-199	-68
2017	-98	-184	86
2018	-82	-228	146
2019	21	-241	262
2020	-166	-334	168

Miscellaneous information such as holiday date was included into the exploration. There are several major vacation periods in Portugal: New Year (1st of January), Carnival (variable, between February and March), Easter (variable, between March and April), and

Christmas (25^{th} of December).

3.3 Approach

The standard software development methodology is not completely suitable for the data analysis programming stage description. As a rule, working on an analytical project results in a different approach. There are several steps that guide the main stages for data science projects. These are diverse in details, although the idea remains the same [17]:

- Description of the data set
- Condensation of the original information
- Deepening of interpretation and transition to explanation
- Analysis of sequences and spatio-temporal patterns extraction
- Development of dashboard instruments

Data processing implies two key actions: cleaning and clarification. The cleaning step is the identification of errors and omissions made during the collection and input of information, as well as example correction. The task is to find "outliers" (incorrectly scored answers of the respondent) and logical violations in the course of the interview (for example, not making a transition). The clarification assumes a description of the distribution of the data according to the essential attributes from the point of view of the objectives and the problem.

In the current work, the preparation, cleaning, normalization, and saving in convenient form were executed in the initial step. In the second step, uni-variate distributions were constructed, measures of central tendency and variation were applied, and a review of the primary regularities was provided. The evaluation was capacious and allowed extracting the earliest dependencies, reducing the number of features needed for analysis, and identifying further directions.

Primary data analysis (descriptive statistics) is a branch of mathematical statistics that studies methods of processing statistical data arrays in order to find generalizing characteristics of the array elements, to build a compact and clear description of data arrays, and to identify patterns that appear in arrays and/or stand out from the main mass of observations [18]. Regularities discovered in primary data analysis must be validated using new statistical techniques to ensure objectivity.

3.4 Summary

This chapter described the main issues regarding the understanding and analysis of the data, including the algorithms, factor and approach. These lay the basis for the effective understanding of the whole process and the visualization requirements.

Chapter 4

Analysis and Results

Overall, the analysis was conducted in JupyterLab, which served as an integrated development environment. JupyterLab is the latest web-based interactive development environment for notebooks, code, and data [19]. Python is the most widely used programming language. Because the library stack used is not extensive, it is provided here. The key library was Pandas, which aims to be the fundamental high-level building block for performing practical and realistic data analysis in Python [20]. For mathematical calculations, the Numpy library was used. Seaborn and Mathplotlib were used to visualize data . The Scikit-Learn tool was utilised to do clustering and dimensionality reduction. The PyTorch framework was utilised for the implementation of VAE. The development instruments were chosen based on relevance, modernity, convenience, and efficiency. The special comparable analysis for tool selection was not performed because these are well-known and well established in the data science community.

4.1 Initial Data Processing

The first step in the analysis process is always data collection. The first and most important step is to determine the sources of the materials. Skipping the specified stage may result in the loss of numerous chances, a “limited picture” and incorrect conclusions. In addition, the sources must adhere to the notion of dependability and be authoritative.

In this context, three main categories were identified: digital, city administration, and health. The city administration provided the water consumption statistics. The website Pordata was used as an extra digital source. The Francisco Manuel dos Santos Foundation (FMSF) project is a database of certified statistics on Portugal, its municipalities and Europe. It is a trustworthy, open archive that collects datasets from various fields. The project utilized the variety of Pordata resources, including geographical coordinates of Braganca parishes, annual precipitation data, and COVID-19 case data.

The following step is to understand the provided data, its types and its structure. The first stage of processing allows us to determine the information's fundamental structure, its tendency, and the subsequent steps necessary to better understand regularities and pattern extraction. Essentially, "NaN" (Not a Number) and duplicate values were eliminated. Figure 4.1 displays the initial version of data structure.

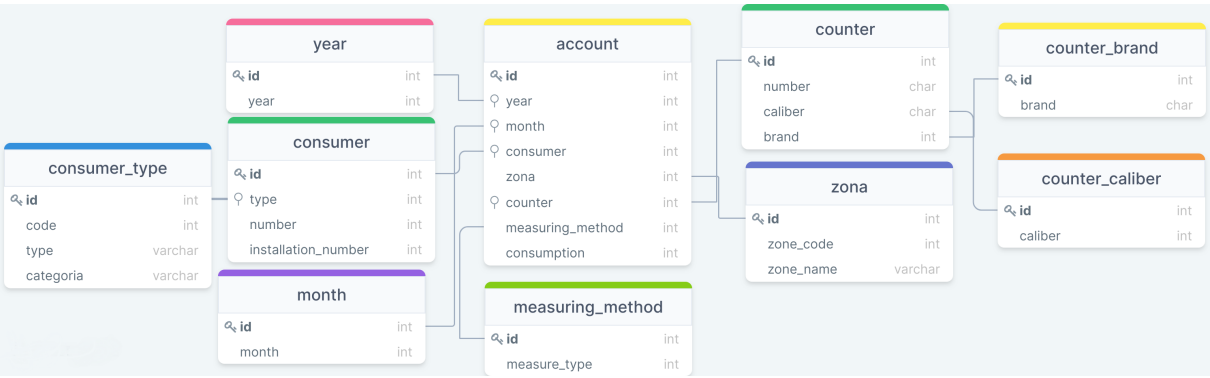

Figure 4.1: Water Consumption Data Structure Scheme version 1

According to Figure 4.1, the primary entity is an account, which is made up of all other entities. A typical water bill includes the following information: id, year, month, data about consumer, geographical zone, information about counter, measuring method and consumption. Consumer data includes individual numbers, consumer type, category and installation number. The installation number should be explained here as the parameter, which is a code for installation location. This is significant because a single consumer may have multiple houses and counters, but each counter has a unique installation number associated with its location. Each counter has a unique number, brand or manufacturer's

name, and caliber. The latter means that the counter diameter, which is determined by the number of taps, determines the size of the counter. Given that the counter caliber has no effect on the amount of water consumed, this parameter can be regarded as secondary.

Further examination of the dataset reveals insignificant parameters that will not be relevant in the analysis. Figure 4.2 shows a simplified scheme with only variables used for exploration.

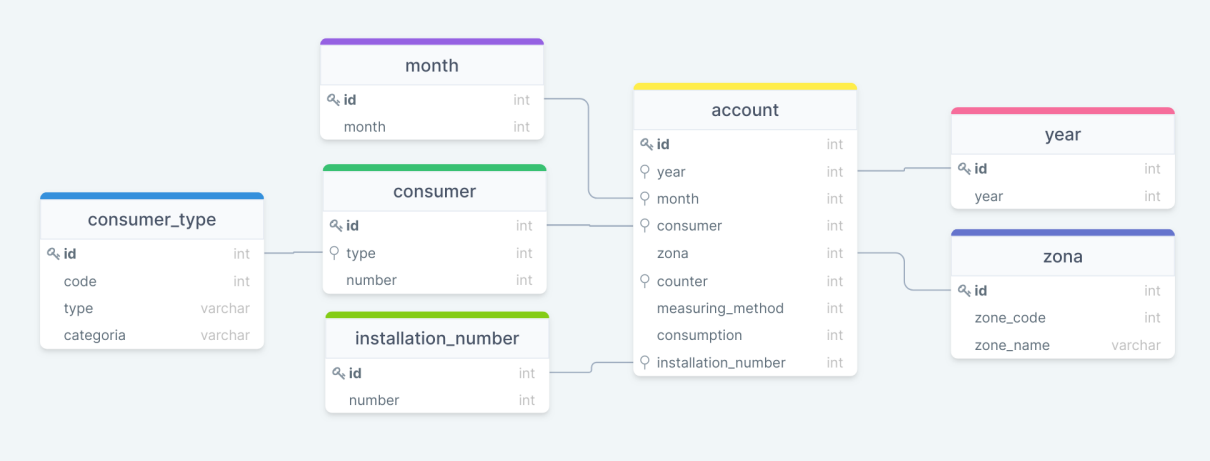

Figure 4.2: Water Consumption Data Structure Scheme version 2

Counter caliber, counter brand, and measuring method have all been removed from the simplified scheme tables. This reduction is possible because the primary goal of the exploration is ”to discover water consumption patterns”. The primary argument in such an analysis is consumption, which is unaffected by counter characteristics or measurement method. The table of installation number has been moved to a new column.

Cleaning is the next step in the processing of datasets. The records may be irrelevant, null, have no numerical value, be incorrect, and repeated. Firstly, all repeated and NaN values were deleted to clean the dataset. Some records had negative values, which were due to prepayments. If a consumer has a water deposit, the consumption will be considered paid and recorded as negative. It is justified to take only absolute negative consumption values. Records were checked to estimate null record consumption for all months and years for each consumer. If all months and years had zero values, this consumer would be

eliminated as a non-contributor. The dataset started with 2171458 rows and ended with 352944 rows after cleaning. For all years and months, a total of 6322 consumers had zero consumption.

The final cleaned dataset was saved in three formats: the original unclean dataset, the final cleaned dataset, and cleaned table based on months. The first entry was used to estimate the impact of null values. The cleaned data served as the primary foundation for exploration. The pivoted table has been saved for monthly review.

4.2 Descriptive Statistic

The first phase of analysis is always rudimentary analysis, which entails primary inquiry and interpretation of the central schema as well as the application of methods for processing data arrays in order to find generalizing characteristics of the elements, build a compact and clear description of data sets, and reveal patterns that appear in arrays or sharply distinguished observations. The goal of the stage is to test hypotheses developed during the research planning stage prior to data collection. The primary description techniques are the creation of histograms, scatter diagrams, and other graphical data representation methods. So, review the final data frame once more and highlight the most important parameters (Table 4.1).

Table 4.1: A final data frame sample

Year	**Month**	**Consumer number**	**Consumer type**	**Installation zone**	**Consumption**	**Installation number**
2019	1	21018	1	4	11	1.0
2019	1	49120	1	6	5	2.0
2019	1	17940	1	6	0	3.0
2019	1	14273	1	6	1	4.0
2019	1	5	1	6	8	5.0

The final version of the records base contains only a few significant columns: year, month, consumer number, consumer type, installation zone, installation number, and consumption. Primary statistics are a common and natural first step in any data analysis.

The following are the most important: the arithmetic mean, maximum, minimum values, and standard deviation. The estimated value in this case is consumption, which has been assigned as a fixed variable to examine the variability of a trait under the influence of any controllable factors or to define key aspects. The first of these is the consumer type. An obvious statement is that water consumption varies according to consumer category. The quantitative estimation of expenditure based on consumer type is shown in Table 4.2.

Table 4.2: Water consumption quantitative analysis by consumer type

Consumer type	**Count**	**Max consumption**	**Min consumption**	**Avg consumption**	**Std consumption**
DOMESTICO	244607.0	2119.0	0.0	6.796	8.747
COM/INDUSTRIAL/OBRAS	23889.0	4978.0	0.0	11.066	61.204
OBRAS	2887.0	635.0	0.0	7.483	22.125
UTIL.PUBLICA	885.0	4408.0	0.0	171.325	466.593
ESTADO	1208.0	2710.0	0.0	110.406	294.363
DOM./RURAL A.S	6996.0	1037.0	0.0	5.984	14.476
FAM.CARENCIADAS	1456.0	113.0	0.0	8.810	8.085
CP.DOM/URB	1177.0	97.0	0.0	5.186	7.694
RURAL DOMESTICO	66913.0	920.0	0.0	4.911	8.217
COM./RURAL A.S	177.0	111.0	0.0	14.537	19.211
RURAL/ESTADO	685.0	3611.0	0.0	32.392	248.17
IPSS/IGR/RURAL	286.0	692.0	0.0	59.122	106.593
EXP.A.RURAL	1329.0	100.0	0.0	3.428	7.718
CP.DOM/RURAL	2.0	0.0	0.0	0.0	0.0
FAM.NUMEROSAS	73.0	52.0	0.0	20.096	10.544
CP.COM/RURAL	17.0	24.0	0.0	3.882	6.508
IGREJAS	47.0	29.0	0.0	4.085	7.371
REGA	140.0	3298.0	0.0	275.257	618.51
CMB	170.0	2115.0	0.0	87.859	239.445

Overall, the largest contribution of domestic and rural consumers is the lowest account for CP.DOM/RURAL, but maximal water flow is observed for industries and public utilization. Surprisingly, REGA, or irrigation, has the highest average water usage. This can be explained by the fact that the average consumption for watering is greater than the average household need. The irrigation consumer type has the highest standard deviation,

indicating how much variation exists in the value. The CP.DOM/RURAL consumer type, which denotes the separate categories of domestic and rural households, has the lowest characteristics.

The second step is to determine how consumption changes depending on the geographical zone. According to the above analysis of different consumers, the areas with the greatest amount of public infrastructure, factories, and significant cultivated land are crucial. Table 4.3 summarizes water usage in each territory in the Braganca region.

Table 4.3: Water consumption quantitative analysis by installation zone

Zone	**Count**	**Max consumption**	**Min consumption**	**Avg consumption**	**Std consumption**
Samil	81821.0	3437.0	0.0	7.798	29.158
Santa-Maria	74048.0	2710.0	0.0	8.576	45.536
Gimonde	70733.0	4978.0	0.0	8.061	32.969
Sé	53918.0	4408.0	0.0	9.095	63.689
Izeda	4681.0	3611.0	0.0	10.069	96.563
Pombares	255.0	22.0	0.0	3.752	3.756

The parishes with the most consumers are: Samil, Santa-Maria, Gimonde, and Sé. Samil includes one primary school, one church, one hotel, 28 auto salons and car repair shops, and one large Burger King restaurant. This zone is defined as a region with densely packed public spaces, which explains the high water flow. Santa Maria is the city center of Braganca. There are museums, tourist attractions, supermarkets, markets, shops, and administrative buildings. The high water consumption is typical for this area. Gimonde is an agricultural sector, with many domestic and rural households. Farms typically use a lot of water for irrigation or for cattle. Sé is the last region to contribute the most to water consumption. This area, like Santa-Maria, is known as the central, and it consists of shops and houses. It houses Braganca's largest hospital as well as a secondary school. The water absorption distribution corresponds to expected values. This analysis allows us to conclude that the data is reliable, the cleaning process was not disruptive, the data set under consideration is correct, and the principal consumption relates to the central

regions.

Not all 49 zones were listed in Table 4.3. The regions with the highest and lowest contributions were chosen. The earlier analysis is confirmed by the geographical area-specific description of water allocation. Gimonde has the highest consumption, which can be explained by large watering territories. The Pombares region provides the minimum input as well as the rest of the indicators. It is a small territory on the south of Braganca without irrigated land or urban structures. In 2001 the population of this area was 59 people.

The primary focus following data identification is the central tendency in consumption by years. Actually, it is important to know whether water consumption increases or decreases with age in the area in question. Figure 4.3 demonstrates the total amount of water a consumed from 2013 to 2020.

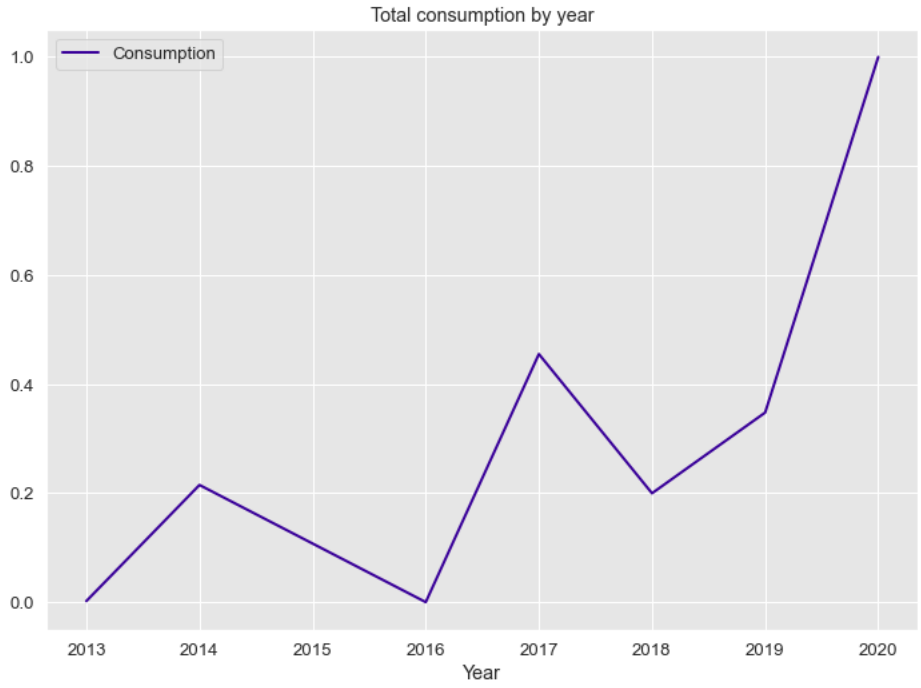

Figure 4.3: Total water consumption annually

The trend of rising consumption is being tracked over time. The sharp rise was not observed in the years preceding 2018, but changes in general can be described as slightly

increasing. It began to rise dramatically in 2018 after a minor decrease and gradually increased until 2020 with no significant fluctuations. Presumably, there are several possible explanations for these non-proportional deviations in resource usage: growth of population, expansion of farming land, creation of new or enlargement of existing production facilities, and huge events that attract a large number of tourists. Certainly, only one of the individuals listed could play the role, but it could be important all at once or in multiples. It would be derived from the further analysis

First and foremost, changes in consumer behavior are reflected across various consumer types. If a new product is introduced, the contribution of industrial types to total consumption should increase, and if the population grows, the influence of the category "DOMÉSTICO" should increase. Figure 4.4 reflects total water usage by consumer type and aids in tracing the impact of each on shifts in consumption.

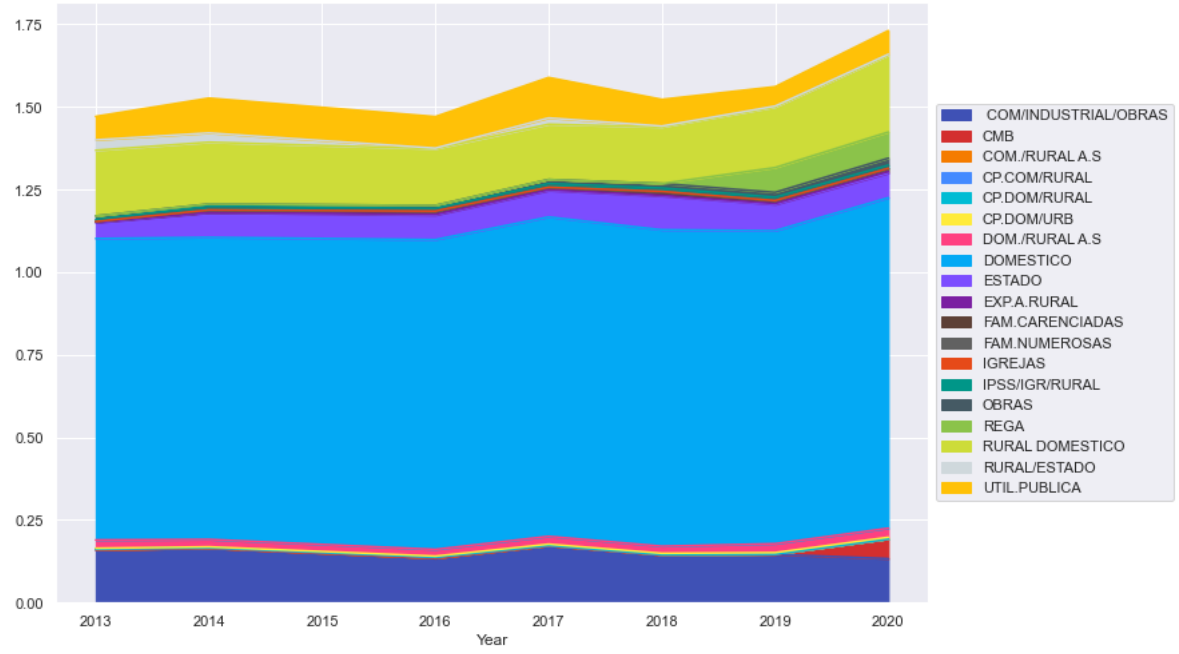

Figure 4.4: Water consumption over time for each type of consumer

Domestic and rural households account for the majority of spending. Another remarkable fact is that "REGA" water consumption has been increasing since 2018. This may explain the substantial increase in overall consumption in 2018 displayed in Figure 4.3.

Furthermore, ”CMB” is rising in 2019 and contributing to the overall increase. Meanwhile, the manufacturing sector is experiencing a slight decline in 2019, which could be the result of a pandemic and quarantine in 2020. In general, excluding the last couple of years, the line eventually stabilizes and shows no detectable fluctuations. This situation can be explained by the fact that water for private use remains roughly constant, and significant increases or decreases usually occur in the sequence of high-profile events such as migration. In fact, the irrigation sector and CBM made a truly significant contribution to water intake.

The majority of consumers belong to ”DOMÉSTICO” and ”INDUSTRIAL” categories, as shown in Table **??**. Figure 4.5 illustrates the monthly distribution of the total consumption over the years.

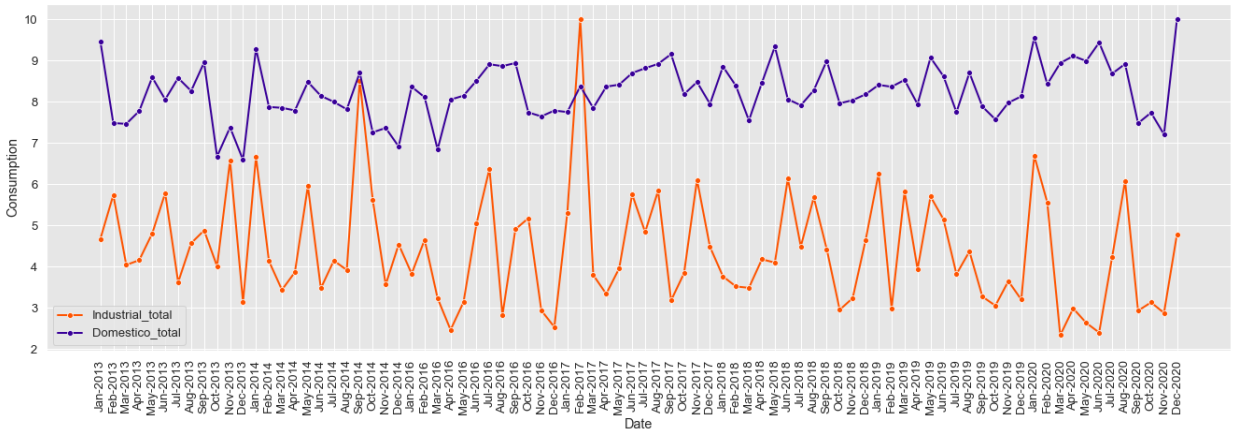

Figure 4.5: Monthly distribution of total consumption for ”Industrial” and ”Domestic” types

The calculation of the monthly water consumption is broadly similar between two types of consumers. Peaks in ’INDUSTRIAL” correspond to peaks in ”DOMÉSTICO”, and falls in the domestic category almost always repeat the falls in industries. The general pattern is repeated across many sites. At this point in the analysis, determining the causes of peaks and falls is difficult, but duplicates of consumption behavior for different consumers are to be expected. It is related to the monthly amount of precipitation and

weather conditions. It does not require a lot of water in cold, rainy months, whereas in hot months, such as summer, people increase their expenditure. Furthermore, lifestyle, traditions and events may have an impact. National holidays are an example, because major celebrations are expected to require more resources. Consider the same chart with precipitation analysis and Portugal feasts. Figure 4.6 demonstrates the monthly distribution of total consumption for "INDUSTRIAL" and "DOMÉSTICO" types with the view of the periods of vacations.

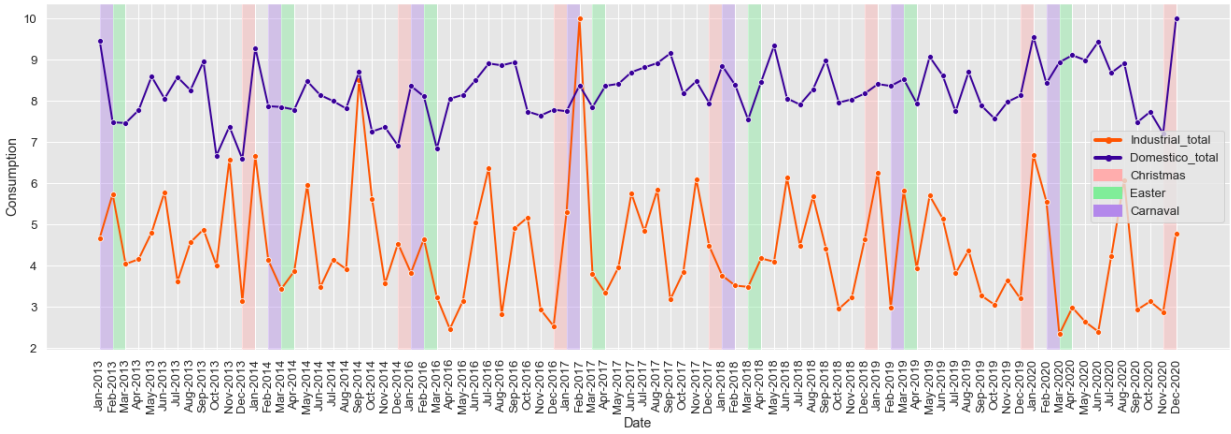

Figure 4.6: Monthly consumption in conjunction with commemorations

In Portugal, there are 13 official holidays, each of which is a day off for workers. As the current study focuses on monthly periods, feasts with extensive commemoration, long preparation period, and home celebrations are of particular interest. For example, if the Christmas holiday lasts for three or four days, followed by New Year's, planning typically begins several weeks in advance. As a result, Christmas, Easter and Carnaval are selected. Carnaval consists primarily of outdoor activities, but the preparation and scope of the event imply a significant impact on water consumption. According to Figure 4.6 overall expenditure in the month of Carnaval is decreasing for domestics and industries, with only 2019 showing growth. Both categories of consumers are experiencing expectedly stable increases for Christmas and New Year. The Easter period is not unambiguous; in some years water consumption remains constant, while in others, it rises and falls.

Easter celebrations are unlikely to have a statistically significant impact on consumption. Another factor that clarifies the chart results is the precipitation level. The logic is as follows: more drizzle leads to less intake, and vice versa (Figure 4.7).

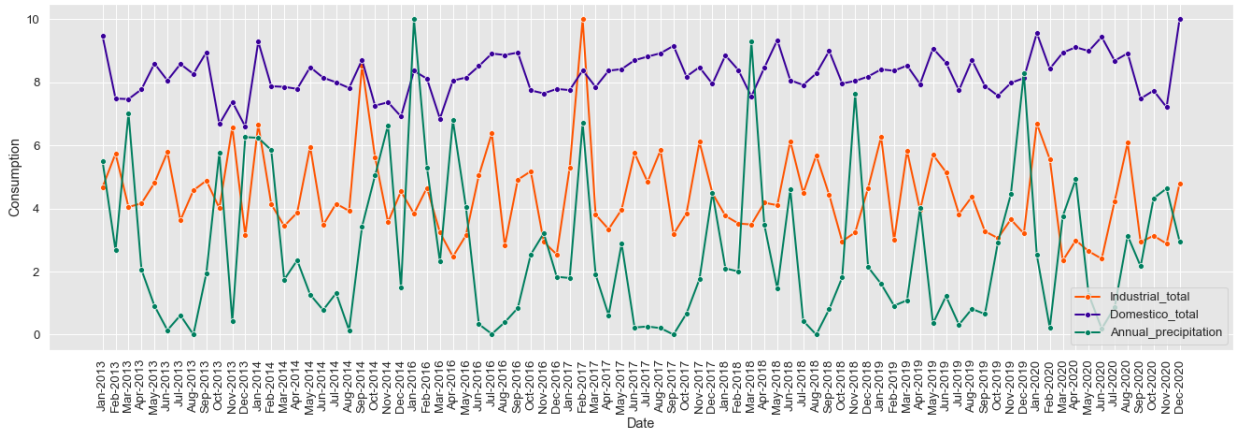

Figure 4.7: Monthly total consumption for each year based on precipitation level

The first peak of rain corresponds to the drop in water intake; as the amount of rain was fell steadily, consumption increased. The described fluctuation has a yearly occurrence. The periods with the highest fall outs are from November to April, while the periods with the lowest grade of rain from June to October. Summer water consumption is significantly less than during the winter. However, exceptional behavior was observed in 2017. The maximum precipitation occurred in February of this year, as did the peaks of domestic and industrial water expenditures. Return for a second look. Figure 4.6 shows that this peak corresponds to the Carnaval commemoration. Because the Carnaval fell on February 28th in 2017, it is reasonable to assume that the celebration was larger than in previous years. This expectation is confirmed by information about the festival's organization provided by the municipality of Braganca. It was the largest commemoration in living memory, according to it. Thousands of people from Portugal and Spain took part in street activities. In conclusion, the 2017 values are not exceptional and have compelling justification. The trend of decreasing water usage during periods of low precipitation and increasing it during periods of high precipitation continues.

Population is another criterion for water intake evaluation. Obviously, increased population growth leads to increased expenditure, so the reason for the dramatic rise in total consumption might be migration or sharp increase in birth rates. Picture 4.8 illustrates the dynamic of population changes for the given years, as well as a chart of total water consumption.

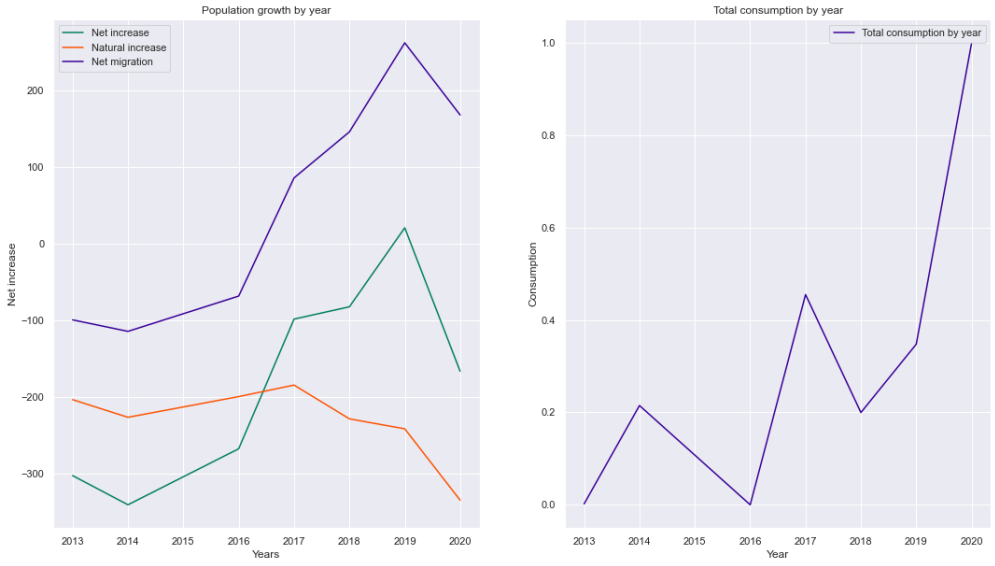

Figure 4.8: Population dynamic and total water consumption by years.

Three major gauges were selected for the analysis of the demographic situation in the Braganca region: net increase, natural increase, and net migration. The net increase is the overall difference in the number of people over time. The difference between the number of people in a given year and the previous year is taken into account. The natural increase is the change caused by natural birth and mortality rates. Only those born or died in the area are counted. Net migration is the criterion for population growth that is defined by the number of immigrants arriving or entering and emigrants leaving or departing. The difference between these two figures is referred to as net migration. Considering the dramatic increase in consumption for the 2020 year, the demographic situation from 2019

to 2020 is of particular interest. During the given time period, there is a rapid decrease in population but a significant increase in water expenditure. The population decline is linked to the start of the Covid-19 Pandemic in 2019.

This raises the legitimate question of whether the increase in water consumption is related to the pandemic. The precipitation analysis explains the dependency between water usage and rains, but it does not explain the dramatic rise in consumption in 2019. The pandemic and quarantine may be the true causes of these unusual shifts. During the 2020 lockdown, the majority of the working population stayed at home. In given situation natural resource use is growing for the following natural reasons.

Firstly, people who stay at home start cooking more. Students and schoolchildren typically prefer to eat at the canteen, while employees frequently choose to eat at the cafeteria for lunch and breakfast because it is less time-consuming and more convenient. Water costs can not be the same to prepare individually and in public areas, and the latter must be significantly less because when one person cooks for ten people, it is different when ten people prepare food for themselves, due to differences in cooking methods, taste preferences and the cost of food in restaurants. This is self-evident and does not require a massive statistical basis. Secondly, strengthening hygiene measures during the pandemic force people to use more water. Thirdly, frequent bathing due to extra time available, increased drinking, and increased cleaning cause the growth in expenditures.

The coronavirus pandemic must have had a serious impact on consumption rise falling in the year 2020. Figure 4.9 shows the graph of COVID-19 cases in 2020 and Figure 4.10 illustrates the lockdown periods.

The charts in Figure 4.9 are quite unanticipated. Water usage fluctuates slightly and is not consistent with any of the previously considered factors. The chart rises slowly during the first lockdown and peaks only in the middle of the second quarantine, with a dramatic increase only in December, which can be connected with previously negotiated holidays. There is no correlation between positive COVID-19 cases and consumption. Probably, it is necessary to consider the water data for the last two years and take into account 2019 for assessment in comparison. It is convenient to make the order of values

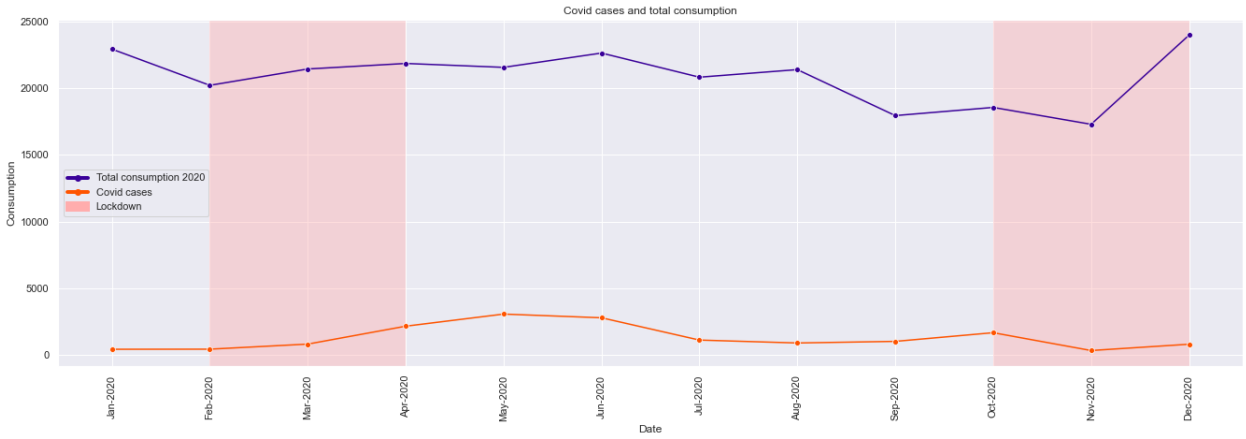

Figure 4.9: Total 2020 water consumption, COVID-19 cases, and lockdown periods

the same and add normalization to improve visualization. Figure 4.10 contains COVID cases, lockdown data, and domestic total water consumption for 2019, 2020 years with normalization of COVID-19 cases data.

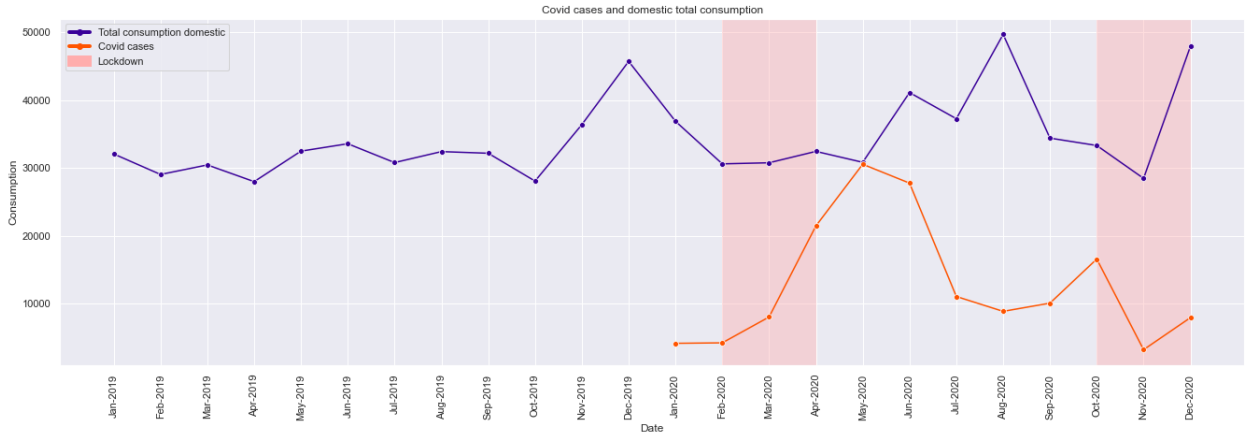

Figure 4.10: Domestic total water consumption for 2019-2020, Covid-19 cases and lockdown periods

The comparison with 2019 improves the picture and highlights the higher volume of expenditures in 2020. However, the tendency and lack of correlation remain permanent, and unfortunately, the pandemic does not explain water usage behavior.

At the moment, the base analysis has been completely finished. The main data set

structure was examined, major tendencies and dependencies have been described. Further study in the field of primary methods results in the accumulation of redundant and ineffective information. The next step of a more complicated examination will be provided in the following chapters.

4.3 Exploratory Data Analysis

The key problem of data analysis is identifying valuable information which can be extracted from large amounts of examples. A brief review of the provided datasets and the preliminary investigation helped in the understanding of data structure and several key regularities. In this regard, it became clear that water consumption increased dramatically in 2018, which was caused more by the expansion of irrigated areas and CMB rather than by pandemic and lock downs. The extent of the water usage strongly correlates with the precipitation levels and can be divided into two periods: summer with the low rate of rainfalls and winter with the sufficient amount of rain. In the summer semester, the amount of water absorbed during the summer semester is significantly greater than during the winter semester. The demographic situation has little influence on water resources and the dramatic decrease in population falls in the sharp increase of water consumption in 2018 is not evidence of inverse dependence; it is simply a fact that indicates that other factors have a greater impact on resource expenditure.

There is still a set of questions to be analyzed, such as how water consumption depends on geographical zone, installation number and consumer type; what patterns emerge from expenditure behavior; how pandemic changes affect resource usage; and how detected characteristics can be used to predict pandemics and other social phenomena?

These opening questions set the stage for further data analysis, though some may go unanswered. As a result, the goal of this section is to perform additional data processing to obtain more appropriate proof of the previous explanations, to extract regularities and to examine the period from 2018 to 2020 in term of details.

Firstly, define the core parameters of the exploratory data analysis. As previously

stated, water usage is determined by the installation number, installation zone, consumer type, and number. Set these variables as the primary search variables.

Table 1 provides a brief description of the chosen criteria 4.4.

Table 4.4: Short description of the parameters.

Parameter	**Unique values**
Installation zone	49
Installation number	21508
Consumer type	19
Consumer number	29692

There are 21508 different installations, 29692 unique consumers, 19 consumer types, and despite the fact that the total number of installation zones should be 66, only 49 zones are encountered across the entire dataset of 352944 records.

The consumer type was chosen in the first parameter's quality to compare the effects on water usage caused by different consumers. To build the yearly consumption, 50 random consumers were chosen, one for each type. During the process of selecting random consumer numbers, it was discovered that only nine categories have consumers complying with the total monthly expenditure for each year that is higher than zero. The results of an operation are illustrated in Figure 4.11.

As the results are irregular, there are no patterns or dependencies in user behavior. There are some months with plenty of zero consumption values. The absence of any correlations between water usage of different independent consumers can testify to a wrong analysis method or implementation errors. The final point can be excluded by considering multiple confirmations of data validity from previous examinations. Further research can verify the exploration approach. If it is incorrect, some beneficial effects for other techniques or in one of the possible scenarios where there are no patterns in water using style should be identified.

All of the statements above attempt to examine associations for the same consumer in different years, taking several chance users from various categories and comparing water expenditure over the course of seven years. Figure 4.12 shows the results for fixed

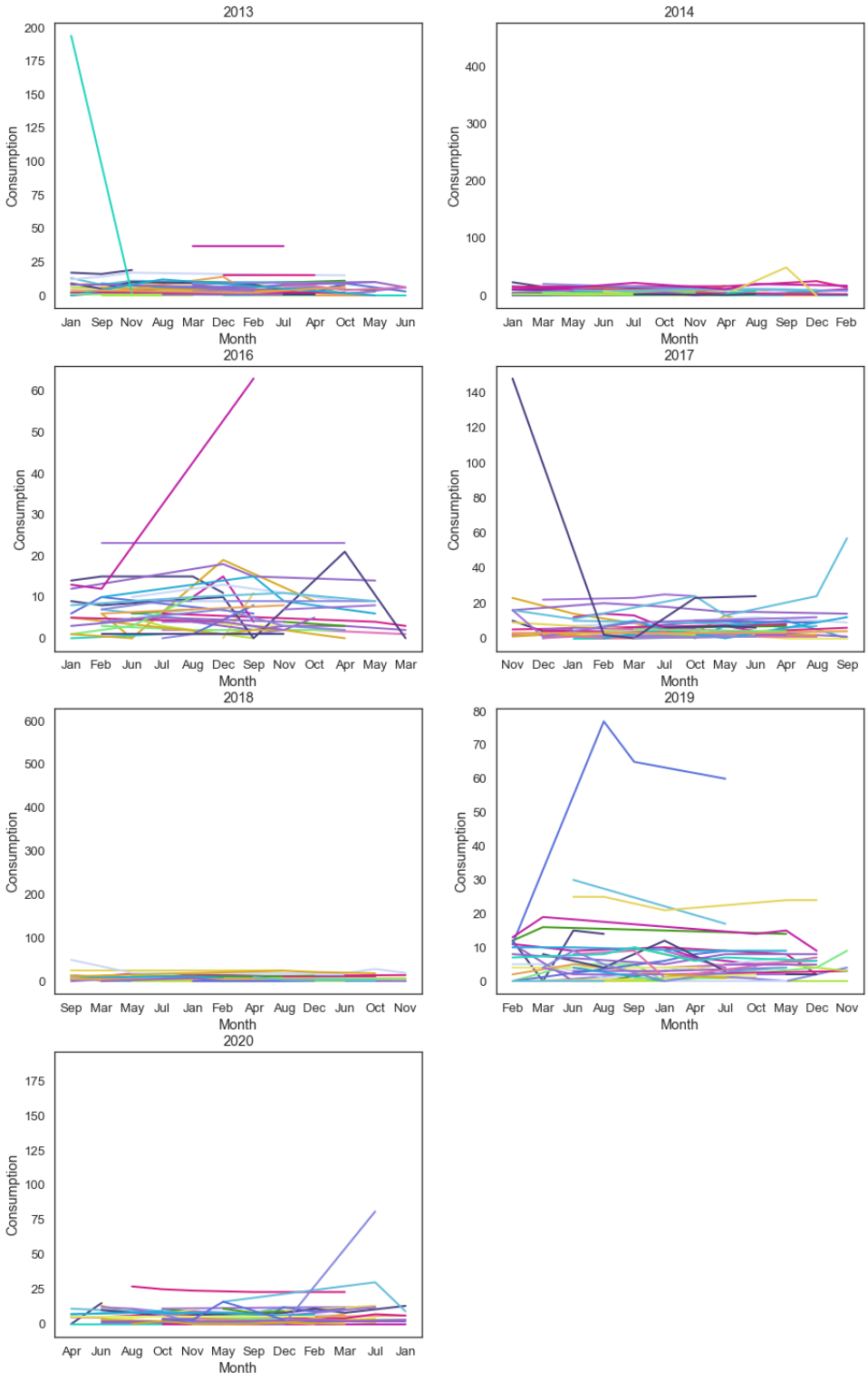

Figure 4.11: Monthly total consumption for 50 random consumers

consumers.

Only arbitrary consumers from the most valuable categories were taken into account. The types with the greatest contribution are implied in Table 4.5: DOMÉSTICO, RURAL DOMÉSTICO, COM/INDUSTRIAL/OBRAS, UTIL.PUBLICA, ESTADO.

Table 4.5: Total consumption by consumer types

Consumer type	**Consumption**
DOMÉSTICO	1662340
RURAL DOMÉSTICO	328654
COM/INDUSTRIAL/OBRAS	264360
UTIL.PUBLICA	151623
ESTADO	133371
DOM./RURAL A.S	41862
REGA	38536
RURAL/ESTADO	22189
OBRAS	21604
IPSS/IGR/RURAL	16909
CMB	14936
FAM.CARENCIADAS	12828
CP.DOM/URB	6104
EXP.A.RURAL	4557
COM./RURAL A.S	2573
FAM.NUMEROSAS	1467
IGREJAS	192
CP.COM/RURAL	66

The greatest number of users does not always imply the greatest total water intake. As a result, absolute values for each type were calculated.

The representation of the water usage over time for different consumers looks better than the representation for random consumers. A pattern can be seen in Figure 4.12 for the types 'DOMÉSTICO" and 'RURAL DOMÉSTICO". The rise in rural areas is lower during the summer,which is likely due to active farming and increased irrigation. Domestic consumers have a concentration of behavioral lines, but if you look closely at the chart, it becomes clear that the trajectory repeats for certain years, but not with the same regularity as in rural cases. The other consumer groups have the same arbitrary distribution of consumption over time.

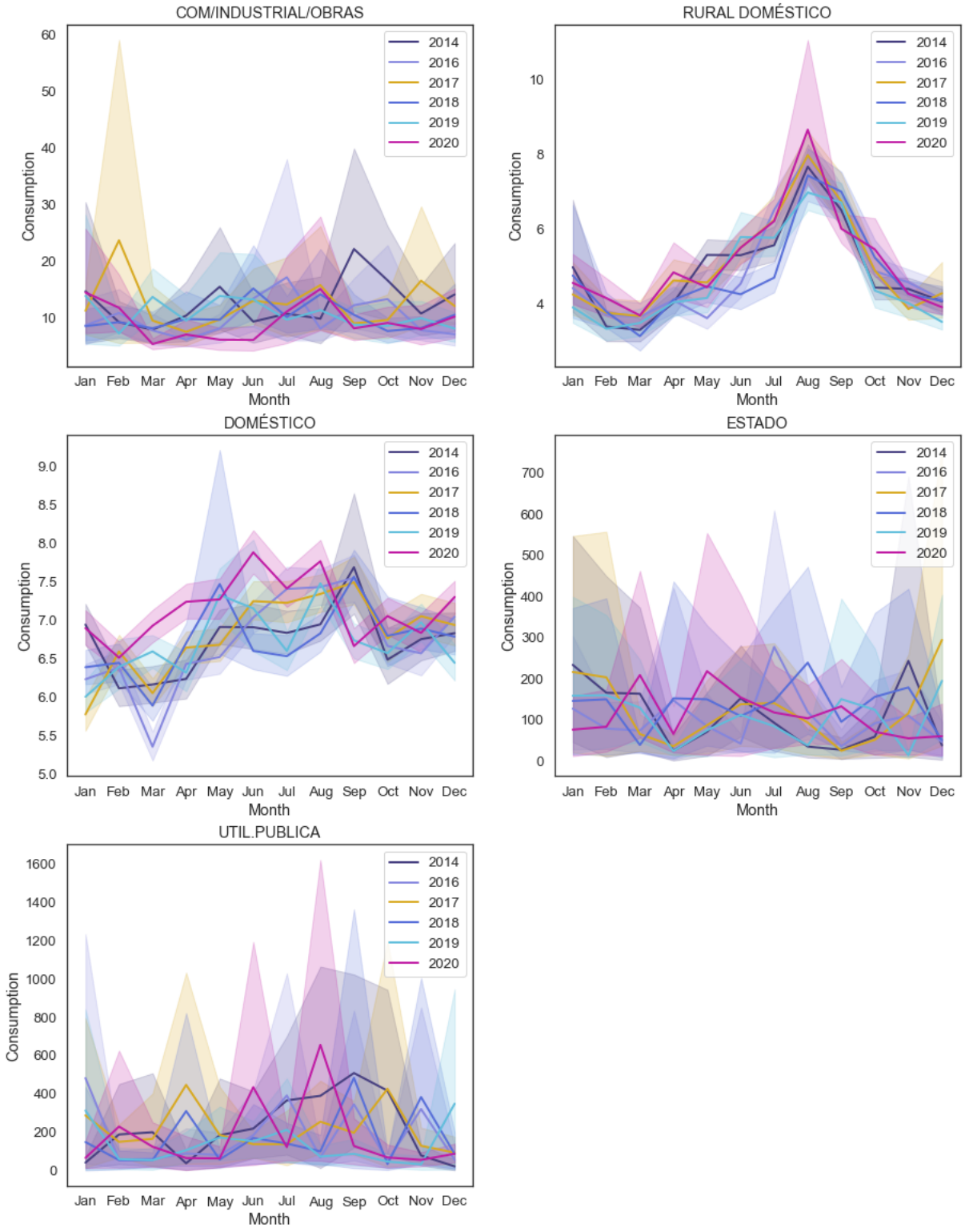

Figure 4.12: Total consumption of different consumer types for all years

The obtained results demonstrate a variety of confirmations of a lack of patterns in water expenditure. A sample of two different parameters gives similar results. This fact allows us to assess the veracity of the processing methodology and the absence of sought regularities. Although conclusions can not be reached at this stage, additional examinations are required.

The relationship summary expenditures from geographical zone has already been slightly affected in the preliminary analysis. The perishes with the most consumers were Samil, Santa Maria, Gimonde, and Sé, but a large number of records does not imply a massive consumption value. The statistics for the geographical zones of the Bragança region are provided in Table 4.6.

As a result, Gimonde and Samil consume the most water, as expected, but Sé and Santa Maria consume the least amount of water. It should be mentioned that the study uses the old congregational division, which saw Sé, Santa Maria, and Meixedo merged into one parish.

The annual dynamic of water utilization in relation to geographical zones is reflected in Figure 4.13. The changes for the specified time interval may be considered fairly steady. The highest consumption is observed for the above mentioned Gimonde and Samil, and Izeda, Calvelhe, and Paradinha Nova. Sé, Santa Maria, and Meixedo continue to have the lowest water intake. The remaining sectors rarely change their values or ignore shifts. However, the general trend of rising expenditures is clearly visible.

A similar trend analysis of water flow rates by territory has been performed for various types of consumers. The heat map was used instead of a geographical representation. Every square on the heat map represents a parish. The purpose was to determine the direction of consumption fluctuations for each consumer type based on zone and area distribution of the concrete category. The specified processing is represented in Figure 4.13. In the current study there are 19 types of consumers, but only 9 are defined: DOMÉSTICO, INDUSTRIAL, PÚBLICO, RURAL, ESTADO, RURAL/ESTADO, CMB, DOMÉSTICO/RURAL, and REGA. These categories are crucial when considered together.

As illustrated in Figure 4.14, the predominance of Braganca territory does not include

Table 4.6: Total consumption by zones

Zone	**Consumption**	**Zone**	**Consumption**
Gimonde	1205245	Quintela de Lampaças	7283
Samil	1128454	Serapicos	7010
Izeda, Calvelhe e Paradinha Nova	56474	Babe	6871
Coelhoso	19860	Gondesende	6822
Rebordãos	19839	Gostei	6618
Parada e Faílde	16334	Rebordainhos e Pombares	6602
São Pedro de Sarracenos	15914	Castro de Avelãs	5895
Nogueira	14403	Pinela	5394
São Julião de Palácios e Deilão	12696	Carragosa	5287
Rio Frio e Milhão	11556	Zoio	4757
Grijó de Parada	11200	Sendas	4755
Outeiro	10914	Rabal	4685
Quintanilha	10682	Alfaião	4465
Santa Comba de Rossas	10630	Sé, Santa Maria e Meixedo	3259
Macedo do Mato	10508	Espinhosela	10349
Baçal	10245	Parâmio	9737
Aveleda e Rio de Onor	9409	França	8920
Salsas	8779	Mós	8412
Sortes	8230	Castrelos e Carrazedo	7905
Donai	7773		

all of the consumer categories listed. The high spot is observed in the following areas: Gimonde, Santa Maria, Sé and Samil. These are the main sectors and every consumer group is undoubtedly represented. At the same time, Sé and Santa Maria have the lowest consumption of all sections. Rural ”DOMÉSTICO” and ”RURAL/ESTADO” are the two classes found in almost every parish. It implies private farms in the countryside, and the land belongs to the state. The rural domestic is distributed uniformly throughout the area, though extremes are already being observed in several territories for rural states.

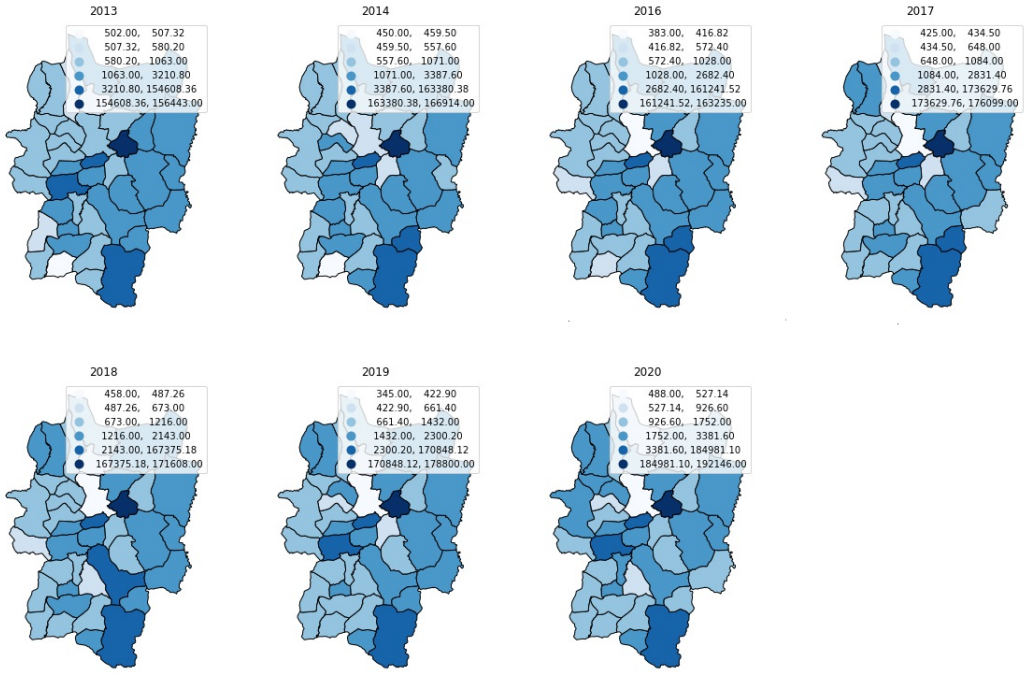

Figure 4.13: The consumption distribution by Braganca geographical zones

Both types exclude central regions such as Gimonde or Samil, which is certainly logical as the area is urban rather than rural. Another odd feature is that the REGA is not spread in the same way as the rural category; backward peaks fall under city jurisdiction, which is likely due to class counts, rather than irrigation of fields or cultivated areas. Generally, the heat maps show that the majority of Braganca is rural land, with the city occupying only five zones.

Briefly, the exploratory analysis was conducted in this section. Here it has been confirmed that there is no correlation between various arbitrary consumers in water usage behavior, weak regularities in consumption curves for rural class over the past seven years have been identified, and for other types of consumers an irregular consumption pattern has been established. The geographically-based water expenditures were analyzed. Identified are the regions with with highest and lowest consumption. The geographic distributions by consumer type and year were provided.

Regarding the preceding analysis, it can be stated that the second stage of processing

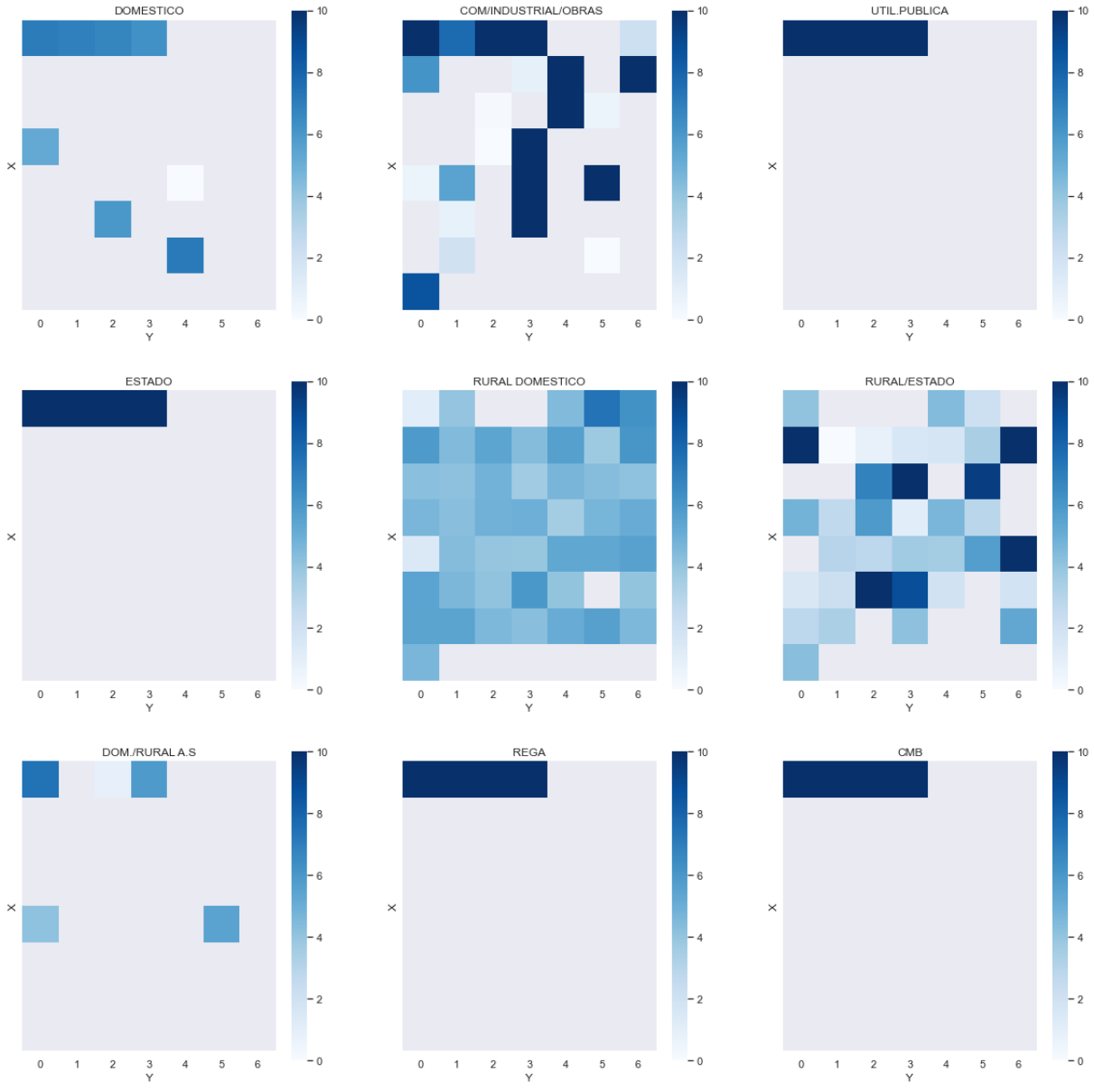

Figure 4.14: Heatmaps of consumption distribution by consumer type

did not reveal any significant evidence of consumption profile patterns.

4.4 Density-Based Spatial Clustering

Clustering is he final stage of the analysis, and its purpose is to identify similarities. Cluster analysis enables the classification of multidimensional data and the separation of a set of objects into homogeneous groups. As seen in the previous chapters, attempts are being made to extract spatio-temporal patterns from water consumption curves. Even though all previous analysis indicated the absence of such patterns, it is necessary to use clustering techniques in order to confirm this statement.

There is a group of objects, consumers in this instance, without defined classes. After clustering the data frame should be divided on classes in such away than inside each bunch there should be "similar" objects, and the objects of separated clusters should be as different as possible.The author suppose that's exactly the kind of manipulation conducted with regard to various characteristics allows to judge conclusively about about the presence of regular structures in the dataset under consideration.

There are many diverse clustering algorithms: connectivity-based, centroid-based, distribution based, density-base, grid-based and others. For this work, the DBSCAN was selected. The algorithm assumes that clustering results can be divided into clusters based on the accuracy of the sampling distribution and considers the relationship between samples in terms of density, continuously extends clusters based on the associations to obtain final results. The basic concept of the algorithm is to find areas of high density that are separated from each other by areas of low density.

The method has a number of advantages, namely: the algorithm is not sensitive to outliers, that is, in the process of clustering all the outliers are taken into a separate cluster with a predetermined label; this method does not require apriori setting the number of clusters; using this method allows to work with clusters of different nature (shape); application of this algorithm allows to work with samples of large volume.

In the present study the clustering has been executed on two counts: installation number and consumer number. Firstly, the consumer number was used, as unique identifier of the individual user. The clustered parameter was consumption, which was examined on a monthly basis. The prepared data frame is provided in Table 4.7.

Table 4.7: Frame prepared for clustering example

Consumer number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	0.0	0.0	0.0	31.0	208.0	245.0	412.0	308.0	0.0	0.0	0.0	0.0
5	8.0	0.0	0.0	0.0	8.0	8.0	8.0	0.0	0.0	0.0	8.0	59.0
11	5.0	0.0	5.0	0.0	0.0	0.0	0.0	12.0	0.0	2.0	0.0	0.0
15	38.0	20.0	13.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	3.0
16	0.0	0.0	0.0	0.0	0.0	0.0	15.0	0.0	12.0	0.0	4.0	0.0

In the given table the columns are months and the values in cells are the water consumption. Therefore there are 12 dimensional data frame. It is necessary to reduce dimensionality. For this, three different approaches were selected: PCA, VAE and t-SNE. We will not dwell on these methods in details, only note that three diverse approaches were utilized to receive a more reliable picture. Also an important fact is that only the last two years, 2019 and 2020, were processed. Figure 4.15 illustrates the DBSCAN with PCA.

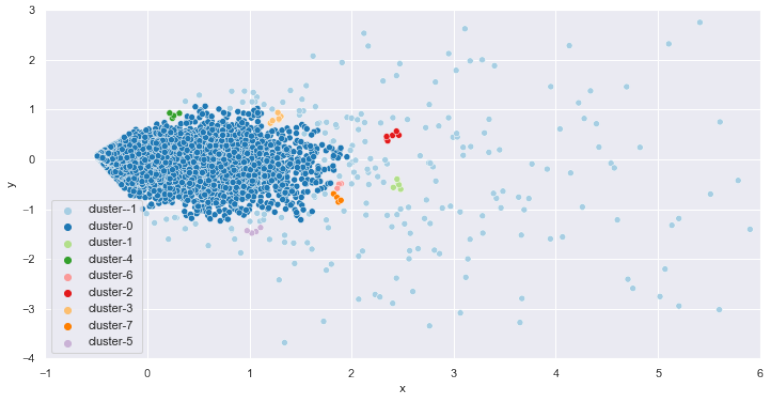

Figure 4.15: DBSCAN clustering with PCA dimension reduce method

The clustering specification is: minimum amount of samples equal 4 and epsilon, the distance between items is 0.1. Mostly these parameters were selected by experience, but few conditions were taken into account. For instance there is no sense to consider groups consisting of 3 points, consumption pattern of 3 different consumers may count by coincidence or statistical error. The distance between points should not be remarkably high compactness.

Altogether Figure 4.15 proves the hypothesis of the absence the classes with similar consumption profiles. There are two big clusters, the first is the biggest group with identical water expenditure behavior, the second most apparently is the onsumers who can not be categorized in any other group. The six little clusters with four or less than ten items merely indicate that small groups of people have similarities in water utilization. Nevertheless at the same time, the number of these classes amends with algorithm coefficients variations, which does not allow to conclude about stability of any particular group and presence of correlation in water expenditure outlines.

With regard to specified investigation it is interesting to differentiate which consumers types make up the obtained clusters. Figure 4.16 illustrates the representation of classes by consumer type.

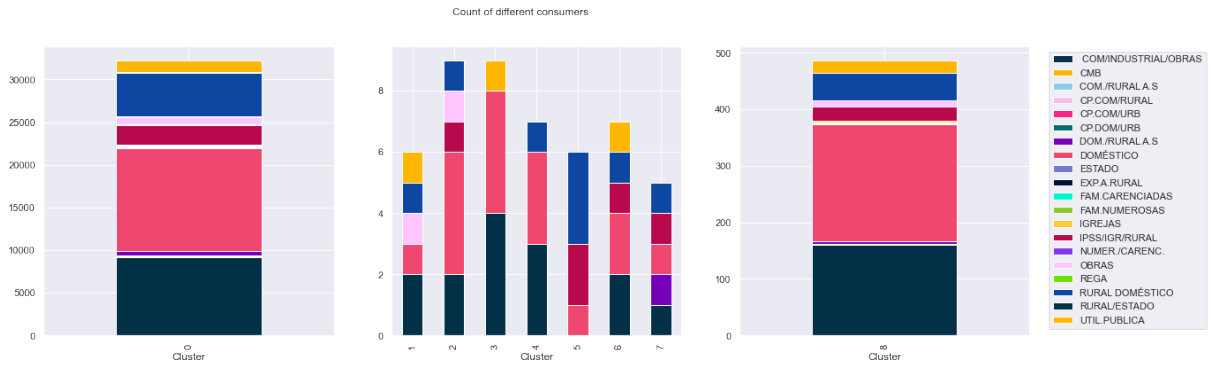

Figure 4.16: Distribution of different consumers for clusters

The allocation by types is quite similar for diverse consumer categories. It indicates that the obtained classes have equivalent profiles according to consumer types. Presented results are already serious evidence on the basis of which it can be concluded the absence of

any permanent regularities in water consumption whose changes allows to predict certain events.

Perhaps there are factors led to the improper clustering operation, wrong dimensional reduce method or erroneous selection of cluster parameter. Figure 4.17 presents the same clustering operation, to reduce number of input variables the VAE was used.

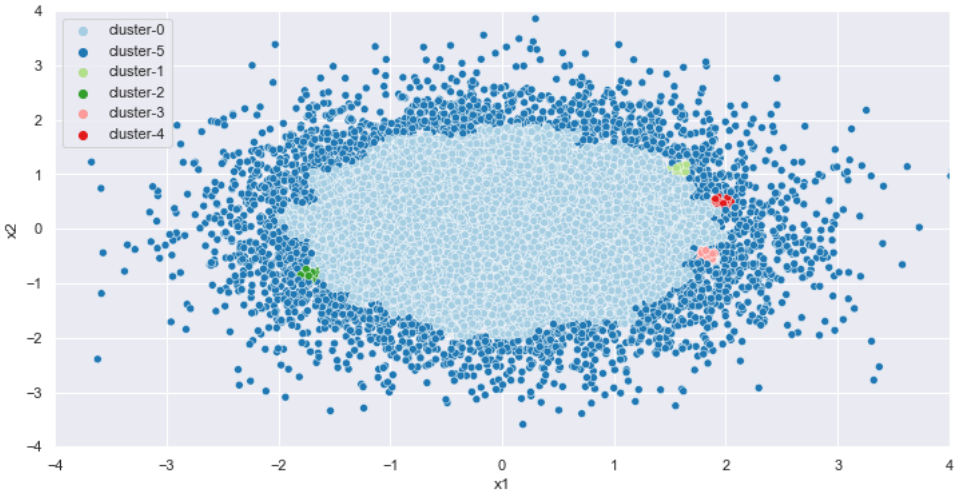

Figure 4.17: DBSCAN clustering with VAE

The VAE managed with task a little better, although the gist remains equal. There are two main classes and random little groups are observed. Similar to the previous outcomes the first class means the points or consumers which have not identical, but enough similar consumption pattern and the second kind is the group of items that just can not be included into another cluster. The little subgroups are arbitrary again and strongly depends on parameters of algorithm.

The profiles in relation to consumer types looks almost equal for different clusters. The scheme is following: the major part of consumers is domestic, the second is the group of rural category and the third largest bunch falls on industrial category. The remained distribution is separated between churches, public utilization and others negligible consumer types. In effect this profile is the typical allocation of users according type, which was demonstrated in figure 4.4 in basis analysis chapter. This proves it once again that

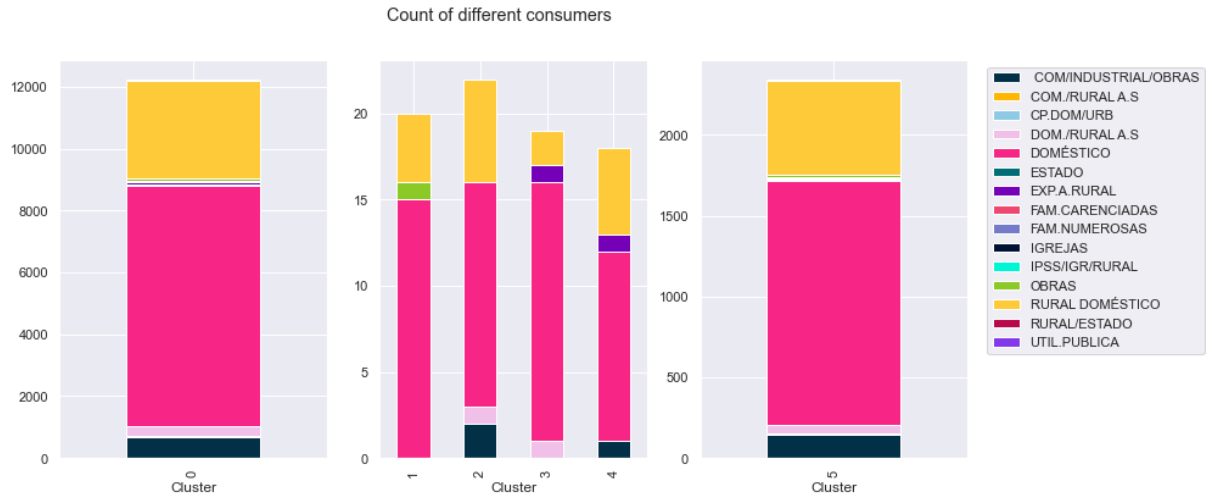

Figure 4.18: DBSCAN clustering with VAE, profiles for different clusters

clustering produced chaotically and there is no any spatio-temporal patterns of water consumption.

The last approach to dimensionality reduction is t-SNE and the results of clustering for it presented in Figure 4.19 and Figure 4.20 respectively.

As can be seen the tendency is retained, the clusters are not accurate, the separation is not not explicit, the profiles are quite identical for clusters. There are a little not checked conjectures, the first is it is necessary to change clustering parameter to installation number instead of consumer number, the second is the distribution is 3-dimensional and the reduction should be produced to three points.

Probably that patterns in water intake does not relies on consumer number and installation number is better indicator. In the process of examine the clustering with installation number, it became clear the results for new parameter is remarkably similar with the above analysis. The whole equal investigation was conducted as for consumer number factor, all three types of dimensionality reduction were applied, but all findings repeated the results for consumer type. It has sense to provide just one clustering type as proof of insolvency. The better results were obtained for t-SNE dimensionality reduction and illustrated in Figure 4.21.

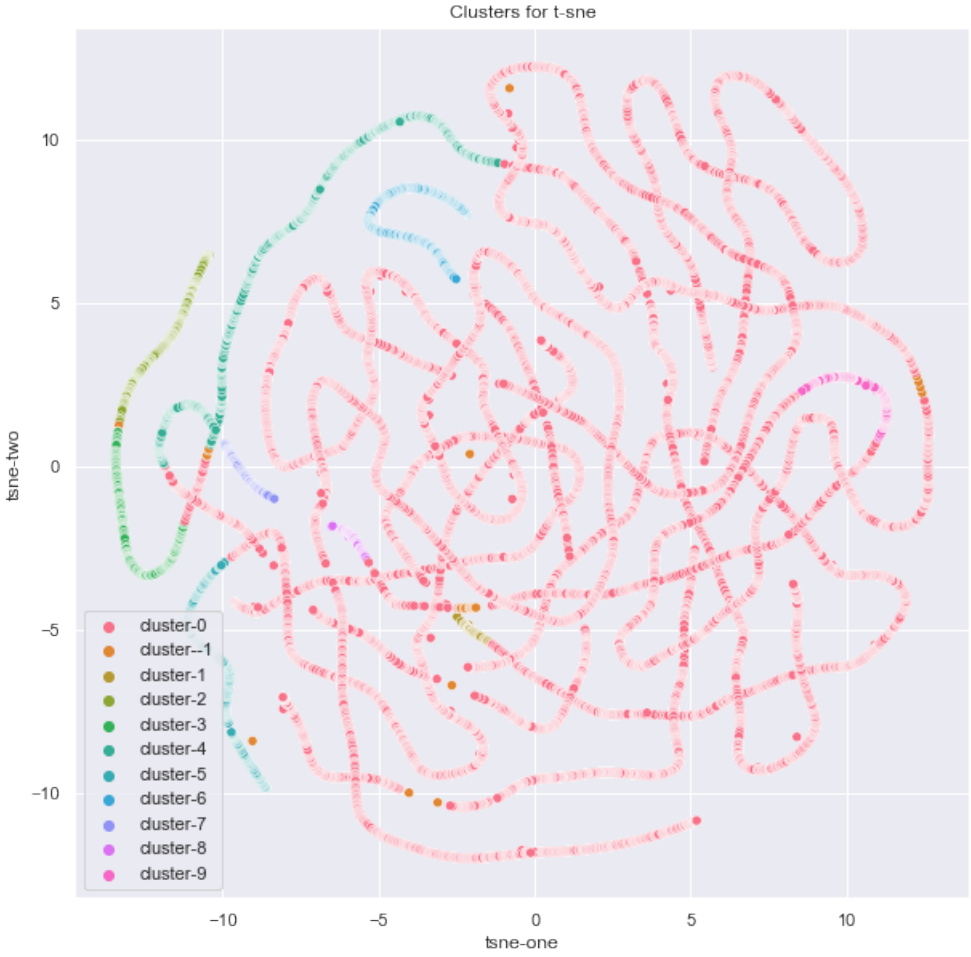

Figure 4.19: DBSCAN clustering with t-SNE

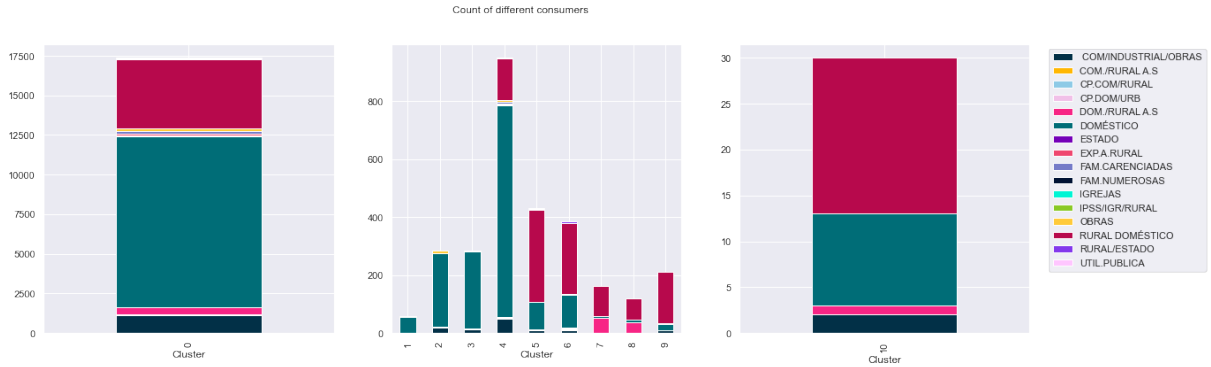

Figure 4.20: Profiles for different clusters

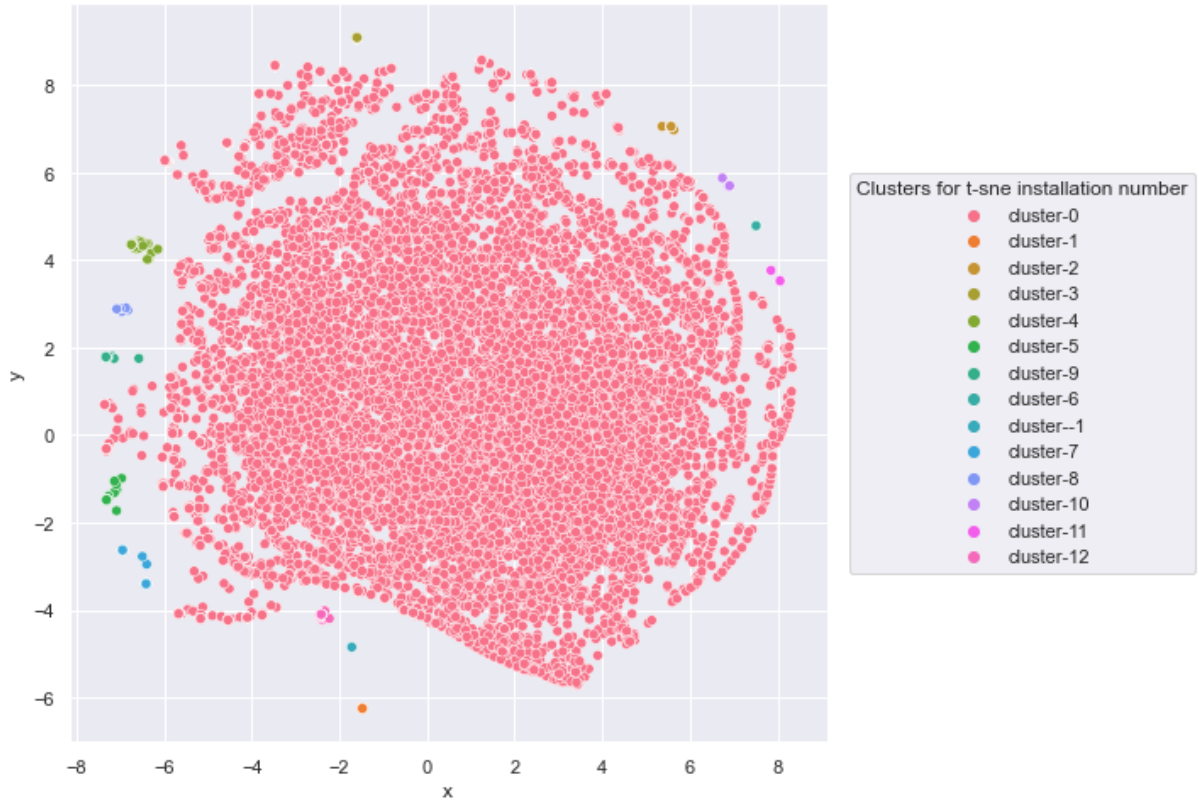

Figure 4.21: DBSCAN clustering with t-SNE by installation number

The profiles of consumer types for installation number criteria replicate the base distribution scheme by consumer types and shown in Figure 4.22. This is the penultimate fact of nonexistence the pattern desired.

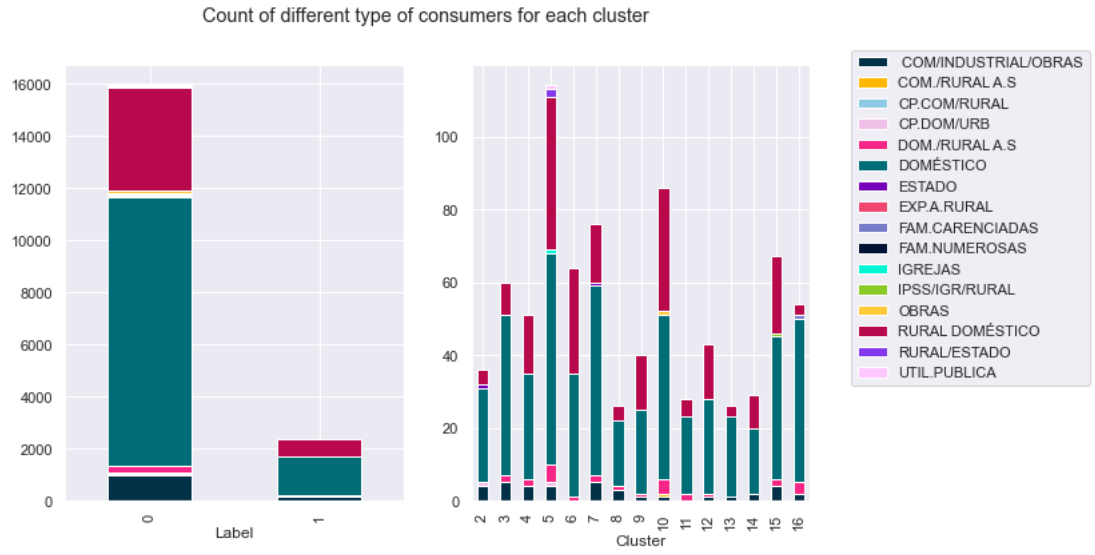

Figure 4.22: DBSCAN clustering by installation number with t-SNE

Already for current status of review it can be concluded that water consumption does not contain any mathematically observable patterns at its core. Nonetheless, it can not be ruled out the guess about the three-dimensional regularities in resources expenditures. As the final of the analysis the 3D chart of clustering produced. The instance of such operation is illustrated in Figure 4.23.

The nature of the data is two-dimensional, hence the spatial distribution is not observed. The processing has generated an equal view for 3D space, a final shape is flat and two-dimensional, meanwhile, there are no clear classes.

Overall the extensive resume is provided in chapter six and a review of the whole offered analysis. Here for summary can be said the detailed cluster processing has shown the absence of spatio-temporal regularities or patterns in water consumption data. Such parameters as individual consumers numbers, various installation zones and consumer types were handled. None of these criteria have demonstrated a statistically significant

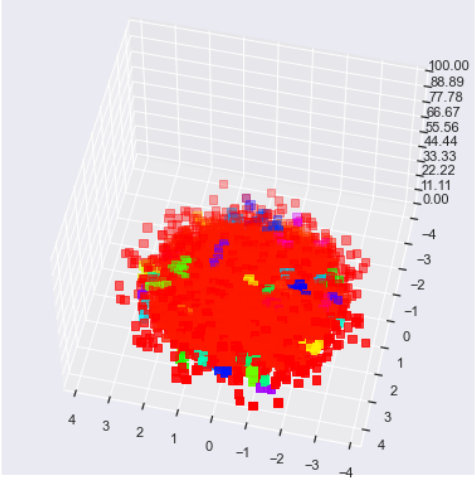

Figure 4.23: 3D DBSCAN clustering with by consumer number VAE

dependency.

4.5 Summary

This chapter described the main analysis process, crossing the water consumption data with with pandemic statistics, population records, precipitation data, and main events. Notable trends were identified. The interpretation of results has different limitations. Depending on conditions, area, and purpose, individual perceptions are needed. This exploratory analysis can be characterized by a deeper understanding of the main characteristics and the association between the features than for basic analysis.

Water consumption was examined in relation to geographical zone and consumer type. The random consumers were compared with consumer of different types over the stated period. The results obtained have already been traced to the absence of the patterns. As the techniques for further investigation, the clustering methods were selected. The DBSCAN clustering method has become the instrument of sequence analysis.

Chapter 5

Visualization and Dashboard

In this chapter, the detailed description of the web-application for graphical representation the performed analysis is provided.

The libraries for chart, graphs and plots creating as Seaborn or Mathplotlib are quite good and allows to generate the graphical representation of good quality. There are obvious minuses of using illustrations inside of jupyter notebook: it requires specific skills, it is necessary to relaunch notebook, any changes lead to errors, it includes code, comments and other unnecessary information. In contrast, the visualization instrument should be available, provide good interpretation and easy access without specific skills.

To implement the tool the cloud based framework, designed specifically for data visualization by Apache Software foundation, "Superset" has been chosen. The instrument is the web-application, which allows visualize the data easy and make it available through the Internet. The options provided by Superset: intuitive interface, the set of visualizations, the visualizations tool can be used without writing code, SQL support. It supports the Gunicorn, Nginx, Apache web servers, different databases engines, Memcached and Redis [21].

The PostgreSQL database was selected to store the information and act as the back-end for Superset. There is the sample of advantages for this object-relational database management system: open source, supporting of all necessary functions, easy to control and configure.

The first step in the dashboard development is data source configuration. The database should be created and connected and the data can be uploaded in CSV format to the database (Figure 5.1).

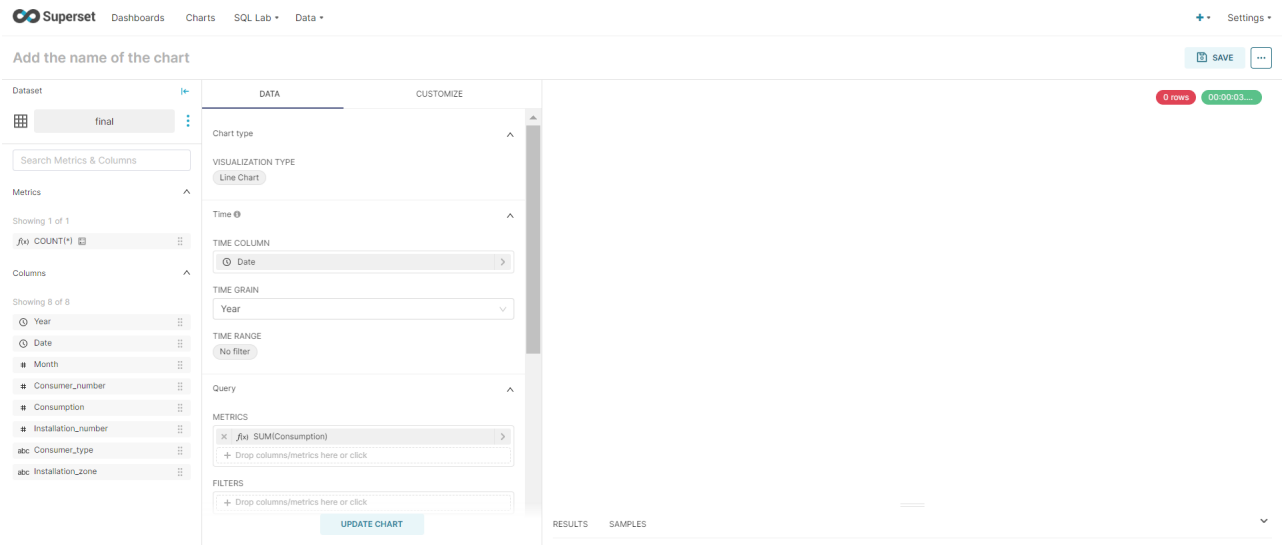

Figure 5.1: Superset charts creating window

The left menu provides the database view, the middle column represents the parameter window for chart generating and the biggest window is for graph representation. The designed chart can be added to dashboard.

There are quite large set of different charts that can be created. The charts are creating by means of SQL requests, which can be created and eddited directly in Superset.

The dashboard based on analysis from chapter 4 have been built. It means that the charts represented in the previous section are available as the web site control panel. Figure 5.2 illustrates the graphical representation of the instrument.

There are dashboard with the various diagrams kit. The charts are dynamic, the value of any point can be checked by the hovering the mouse, the parameters of the plots are editable and can be changed any time for better perception. The order of the sketches can be adapted if it is necessary to compare two or more charts for instance. As well there is a possibility to create tables with different characteristics.

The representation is user-friendly and comfortable, it is simple option to create data

Figure 5.2: Superset dashboard

analytics visualization. Taking into account the fact in the data analysis the decision making remains on researcher or human, the literate and handy visualization is fundamental.

The described step was the last in the thesis, the complete analysis is executed, in the next chapter summarize the work done and results.

Chapter 6

Conclusions

The presented work investigates the spatiotemporal patterns in water consumption and gives the practical base for the further studies in the area of urban data analysis.

The important process of the data collection and preparation was conducted. In the context of city and social intelligence, the influence of pandemic the datasets from various areas were considered: water intakes for the last seven years for Bragança city, demographic statistic, yearly precipitation records, the holidays and big events, the pandemic cases.

The four stage analysis have been performed: data pre-processing, preliminary analysis, exploratory analysis and clustering methods application. The final clauses based on are listed bellow:

- the reason of water consumption dramatically growth in 2018 is the increase of the fields area and irrigation consequently. The proves of other factors contribution was not detected.
- there is the strong correlation between the water usage and precipitation level. Thereby, in the summer period with the low rain frequency, the low level of water intakes has been observed, and the opposite tendency for the winter period, which indicates the strict dependency between these factors.
- meanwhile the connections between expenditures and pandemic cases and lock

downs have not noted. The consumption was not changing significantly during the lockdown periods in comparison with the same periods of previous years.

- the cluster analysis in conjunction with monthly analysis of consumption demonstrates the absence of spatio-temporal patterns in water flow.

- the events impact of water usage, although only large scale. As example from all considered commemorations only the Carnaval of 2017 produced the considerable water intake growth and was the biggest celebration for the whole history. However, it should be noted the stable expenditure increase in the Christmas and new year period.

- the demographic changes for the given period was negligible for creating the statistically significant shifts.

The central purpose of the project was the extraction of spatio-temporal patterns from water expenditures data. Making the whole analysis it can be undoubtedly concluded that this type of data can not be used for predicting models in context of social intelligence. At the same time the results are not useless and can be utilised by municipality authorities for decision making, monitoring and management.

As the additional intention of the thesis the specialized instrument was developed. The tool is a web-application for performing static analysis, including tools to perform slices of intermediate representations of water consumption based on several criteria and tools for visualizing the results of the analysis.

Bibliography

[1] R. P. L. Regina Gubareva, "Big Data trends in the analysis of city resources," *OL2A: International Conference on Optimization, Learning Algorithms and Applications 2022 October, 24 – 25, 2022*, vol. 11, p. 11, May 2022.

[2] O. Kwon, Y. Kim, N. Lee, and Y. Jung, "When Collective Knowledge Meets Crowd Knowledge in a Smart City: A Prediction Method Combining Open Data Keyword Analysis and Case-Based Reasoning," *Journal of Healthcare Engineering*, vol. 2018, 2018. DOI: 10.1155/2018/7391793.

[3] S. Trilles, Ò. Belmonte, S. Schade, and J. Huerta, "A domain-independent methodology to analyze IoT data streams in real-time. A proof of concept implementation for anomaly detection from environmental data," *International Journal of Digital Earth*, vol. 10, no. 1, pp. 103–120, 2017. DOI: 10.1080/17538947.2016.1209583.

[4] M. Bermudez-Edo, P. Barnaghi, and K. Moessner, "Analysing real world data streams with spatio-temporal correlations: Entropy vs. Pearson correlation," *Automation in Construction*, vol. 88, pp. 87–100, 2018. DOI: 10.1016/j.autcon.2017.12.036.

[5] G. Bordogna, A. Cuzzocrea, L. Frigerio, and G. Psaila, "An effective and efficient similarity-matrix-based algorithm for clustering big mobile social data," 2017, pp. 514–521. DOI: 10.1109/ICMLA.2016.188.

[6] G. Wang, W. Wei, J. Jiang, *et al.*, "Application of a long short-term memory neural network: A burgeoning method of deep learning in forecasting HIV incidence in

Guangxi, China," en, *Epidemiology and Infection*, vol. 147, e194, 2019, ISSN: 0950-2688, 1469-4409. DOI: 10.1017/S095026881900075X. [Online]. Available: https://www.cambridge.org/core/product/identifier/S095026881900075X/type/journal_article (visited on 05/27/2020).

[7] R. Pérez-Chacón, J. Luna-Romera, A. Troncoso, F. Martínez-Alvarez, and J. Riquelme, "Big data analytics for discovering electricity consumption patterns in smart cities," *Energies*, vol. 11, no. 3, 2018. DOI: 10.3390/en11030683.

[8] V. Karyotis, K. Tsitseklis, K. Sotiropoulos, and S. Papavassiliou, "Big data clustering via community detection and hyperbolic network embedding in IoT applications," *Sensors (Switzerland)*, vol. 18, no. 4, 2018. DOI: 10.3390/s18041205.

[9] S. Azri, U. Ujang, and A. Abdul Rahman, "Dendrogram clustering for 3D data analytics in smart city," Issue: 4/W9, vol. 42, 2018, pp. 247–253. DOI: 10.5194/isprs-archives-XLII-4-W9-247-2018.

[10] A. Alshami, W. Guo, and G. Pogrebna, "Fuzzy partition technique for clustering Big Urban dataset," 2016, pp. 212–216. DOI: 10.1109/SAI.2016.7555984.

[11] C.-S. Chang, C.-T. Chang, D.-S. Lee, and L.-H. Liou, "K-sets+: A linear-Time clustering algorithm for data points with a sparse similarity measure," 2018, pp. 1–8. DOI: 10.1109/UIC-ATC.2017.8397636.

[12] S. Chae, S. Kwon, and D. Lee, "Predicting Infectious Disease Using Deep Learning and Big Data," en, *International Journal of Environmental Research and Public Health*, vol. 15, no. 8, p. 1596, Jul. 2018, ISSN: 1660-4601. DOI: 10.3390/ijerph15081596. [Online]. Available: http://www.mdpi.com/1660-4601/15/8/1596 (visited on 05/27/2020).

[13] W. Chen, Z. Huang, F. Wu, M. Zhu, H. Guan, and R. Maciejewski, "VAUD: A Visual Analysis Approach for Exploring Spatio-Temporal Urban Data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 9, pp. 2636–2648, Sep. 2018,

Conference Name: IEEE Transactions on Visualization and Computer Graphics, ISSN: 1941-0506. DOI: 10.1109/TVCG.2017.2758362.

[14] B. Simhachalam and G. Ganesan, "Performance comparison of fuzzy and non-fuzzy classification methods," en, *Egyptian Informatics Journal*, vol. 17, no. 2, pp. 183–188, Jul. 2016, ISSN: 1110-8665. DOI: 10.1016/j.eij.2015.10.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110866515000535 (visited on 05/14/2021).

[15] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," en, *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, Jun. 2002, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.122653799. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.122653799 (visited on 11/28/2021).

[16] P. Duggal and S. Paul, "Big Data Analysis: Challenges and Solutions," Dec. 2013.

[17] M. Baillie, S. le Cessie, C. O. Schmidt, L. Lusa, and M. Huebner, "Ten simple rules for initial data analysis," *PLoS Computational Biology*, vol. 18, no. 2, e1009819, Feb. 2022, ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.1009819. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8870512/ (visited on 05/26/2022).

[18] C. B. Thompson, "Descriptive Data Analysis," *Air Medical Journal*, vol. 28, no. 2, pp. 56–59, Mar. 2009, ISSN: 1067-991X. DOI: 10.1016/j.amj.2008.12.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1067991X08002976 (visited on 05/26/2022).

[19] "Project jupyter." (2022), [Online]. Available: https://jupyter.org (visited on 05/30/2022).

[20] "Pandas documentation — pandas 1.4.2 documentation." (2022), [Online]. Available: https://pandas.pydata.org/docs/ (visited on 05/30/2022).

[21] A. S. Foundation. “Superset documentation.” (2022), [Online]. Available: https://superset.apache.org/docs/ (visited on 05/26/2022).