UNIVERSIDAD DE GRANADA

MASTER'S DEGREE IN BUSINESS PROCESS MANAGEMENT AND TECHNOLOGIES

Academic year 2022-2023

MASTER'S THESIS

Exploring Vehicle Behavior in Smart Villages Environments

Author:
Daniel BOLAÑOS MARTÍNEZ

Supervisors:
María BERMÚDEZ EDO
Blanca Luisa DELGADO MÁRQUEZ

DETECCIÓN DEL PLAGIO

La Universidad utiliza el programa **Turnitin Feedback Studio** para comparar la originalidad del trabajo entregado por cada estudiante con millones de recursos electrónicos y detecta aquellas partes del texto copiadas y pegadas. Copiar o plagiar en un TFM es considerado una **Falta Grave**, y puede conllevar la expulsión definitiva de la Universidad.

Exploring Vehicle Behavior in Smart Villages Environments: A Comprehensive Clustering Pipeline for Analysis and Insights

Daniel BOLAÑOS MARTÍNEZ

Palabras clave: IoT, sensores, clustering, pueblos inteligentes, explicabilidad

Resumen

Actualmente, debido al cambio climático y al incremento de turistas en todos los entornos, los gestores de organizaciones ubicadas en parajes vulnerables, y los políticos deben aplicar técnicas de gestión de recursos humanos, materiales y políticas de protección medioambiental, al igual que aplican las empresas en sus gestiones rutinarias, cumpliendo con las leyes medioambientales y siendo proactivos en la aplicación de otras técnicas no obligatorias.

El primer paso para aplicar estas políticas pasa por conocer la situación actual. En este sentido, este TFM aborda el conocimiento del tráfico que circula en un paraje vulnerable, como son los pueblos del valle de Poqueira, situados a los pies del paraje natural de Sierra Nevada. En concreto, hemos realizado un proyecto en el cual se ha diseñado el despliegue de sensores de tráfico, gestionando la implementación con la empresa de sensores y definiendo estrategias para la extracción de conocimiento a partir de los datos recopilados. Posteriormente, se aplican técnicas de aprendizaje automático sobre los datos recogidos para estudiar los patrones de tráfico.

Proponemos un pipeline de análisis de patrones que utiliza múltiples fuentes de datos y simplifica la selección de algoritmos de agrupación y normalización. En nuestra investigación, destacamos la importancia de elegir el algoritmo de normalización adecuado para escalar eficientemente las características de entrada de datos heterogéneos. Esto es fundamental para comprender y analizar los patrones de movilidad. Para validar nuestro enfoque, utilizamos datos de cuatro cámaras de reconocimiento de matrículas (LPR) recopilados durante nueve meses en una comarca de la Alpujarra Granadina. También incorporamos bases de datos adicionales con información sobre origen, ingresos y datos de vacaciones, lo que nos dio un conjunto de datos de más de 50.000 vehículos.

Aplicando nuestro pipeline y analizando este gran conjunto de datos, identificamos diversos patrones de tráfico entre residentes y visitantes en una zona turística rural. Los resultados de nuestro estudio aportan valiosas ideas a los analistas de datos sobre los factores a tener en cuenta a la hora de seleccionar algoritmos adecuados para analizar conjuntos de datos heterogéneos. Nuestros resultados son de suma importancia tanto para los gestores de parques nacionales, que se encuentran en zonas especialmente vulnerables a las amenazas derivadas del cambio climático, como para los gestores de organizaciones ubicadas en estos parajes. Además, proporcionan a los responsables políticos una comprensión más profunda de los patrones de movilidad en áreas sensibles desde el punto de vista medioambiental. Todo ello facilita la consecución de la sostenibilidad de estos territorios en una triple vertiente: económica, social y medioambiental.

Exploring Vehicle Behavior in Smart Villages Environments: A Comprehensive Clustering Pipeline for Analysis and Insights

Daniel BOLAÑOS MARTÍNEZ

Keywords: IoT, sensors, clustering, smart villages, explainability

Abstract

Currently, due to climate change and the increase of tourists in all environments, managers of organizations located in vulnerable places and politicians must apply human resource management techniques, materials and environmental protection policies, just as companies do in their routine management, complying with environmental laws and being proactive in the application of other non-mandatory techniques.

The first step to apply these policies is to know the current situation. In this sense, this master's thesis deals with the knowledge of the traffic that circulates in a vulnerable place, such as the villages of the Poqueira Valley, located at the foot of the natural site of Sierra Nevada. Specifically, we have carried out a project in which we have designed the deployment of traffic sensors, managed the implementation with the sensor company and defined strategies for the extraction of knowledge from the collected data. Subsequently, machine learning techniques are applied on the collected data to study traffic patterns.

We propose a pattern analysis pipeline that utilizes multiple data sources and simplifies the selection of clustering and normalization algorithms. In our research, we highlight the importance of choosing the right normalization algorithm to scale the input features of heterogeneous data efficiently. This is critical for understanding and analyzing mobility patterns. To validate our approach, we use data from four license plate recognition (LPR) cameras collected over nine months in a district of the Alpujarra Granadina. We also incorporated additional databases with information on origin, income, and vacation data, giving us a dataset of more than 50,000 vehicles.

By applying our pipeline and analyzing this large dataset, we identified diverse traffic patterns between residents and visitors in a rural tourist area. The results of our study provide valuable insights to data analysts on factors to consider when selecting suitable algorithms for analyzing heterogeneous datasets. Our findings are of utmost importance for both managers of national parks, which are particularly vulnerable to climate change-related threats, and managers of organizations located in these areas. Furthermore, they provide policymakers with a deeper understanding of mobility patterns in environmentally sensitive areas. This, in turn, facilitates the achievement of sustainability in these territories across three dimensions: economic, social, and environmental.

Yo, **Daniel BOLAÑOS MARTÍNEZ**, estudiante del **Máster en gestión y Tecnologías de Procesos de Negocio de la Universidad de Granada**, con DNI 76592621E, autorizo la ubicación de la siguiente copia de mi Trabajo Fin de Máster en la biblioteca del centro para que pueda ser consultada por las personas que lo deseen.

Fdo: **Daniel BOLAÑOS MARTÍNEZ**

Granada, 20 de junio de 2023

María BERMÚDEZ EDO, Profesora del Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Granada.

Blanca Luisa DELGADO MÁRQUEZ, Profesora del Departamento de Organización de Empresas II de la Universidad de Granada.

Informan:

Que el presente trabajo, titulado **Exploring Vehicle Behavior in Smart Villages Environments, A Comprehensive Clustering Pipeline for Analysis and Insights**, ha sido realizado bajo su supervisión por **Daniel BOLAÑOS MARTÍNEZ**, y autorizamos la defensa de dicho trabajo ante el tribunal que corresponda.

Y para que conste, expiden y firman el presente informe en Granada a 20 de junio de 2023.

Las directoras:

Fdo: **María BERMÚDEZ EDO**
Fdo: **Blanca Luisa DELGADO MÁRQUEZ**

Granada, 20 de junio de 2023

Daniel BOLAÑOS MARTÍNEZ, estudiante del **Máster en Gestión y Tecnologías de Procesos de Negocio de la Universidad de Granada**, declaro explícitamente que el trabajo presentado es original, entendido en el sentido de que todas las fuentes utilizadas se han citado debidamente respetando los derechos de autoría.

Fdo: **Daniel BOLAÑOS MARTÍNEZ**

Granada, 20 de junio de 2023

Acknowledgements

I would like to express my heartfelt gratitude to Maria and Blanca for their invaluable guidance as my tutors. Their expertise and support have been instrumental in the completion of this Master's Thesis. Additionally, I would like to extend my appreciation to Alberto, Jose Luis, and Julian for their exceptional teamwork as members of the Smart Poqueira. Their collaboration and trust have been a significant contribution to this project.

Contents

List of Figures

List of Tables

List of Abbreviations

AEAT	**A**gencia **E**statal de **A**dministración **T**ributaria
ANPR	**A**utomatic **N**umber-**P**late **R**ecognition
API	**A**pplication **P**rogramming **I**nterface
AP	**A**ffinity **P**ropagation
ASC	**A**ttributed **S**pectral **C**lustering
AIC	**A**kaike **I**nformation **C**riterion
BIC	**B**ayesian **I**nformation **C**riterion
CH	**C**alinski-**H**arabasz Score
DB	**D**ata **B**ase
DBI	**D**avies-**B**ouldin **I**ndex
DGT	**D**irección **G**eneral de **T**ráfico
DTW	**D**ynamic **T**ime **W**arping
EM	**E**xpectation **M**aximization
GPS	**G**lobal **P**ositioning **S**ystem
IC	**I**nformation **C**riterion
INE	**I**nstituto **N**acional de **E**stadística
IO	**I**ndoor-**O**utdoor
IP	**I**nternet **P**rotocol
IoT	**I**nternet **O**f **T**hings
LPR	**L**icense **P**late **R**ecognition
MAD	**M**edian **A**bsolute **D**eviation
ML	**M**achine **L**earning
MT	**M**aster's **T**hesis
PCA	**P**rincipal **C**omponent **A**nalysis
SC	**S**ilhouette **C**oefficient
SGTM	**S**ecretaría **G**eneral de **T**ransportes y **M**ovilidad
SSB	**S**um of **S**quared **B**etween
SSW	**S**um of **S**quared **W**ithin

Chapter 1

Introduction

Currently, there are 13.4 billion Internet of Things (IoT) devices. Statista predicts that this figure will increase to 29.4 billion by 2030[1]. These devices form an interconnected network that produces extensive data in numerous social domains. Access to a large volume of data collected by various sensors makes it possible to supervise and manage different aspects of society, including healthcare, evacuation systems, smart environments, and transportation. [2, 9, 25, 15]. Extracting and combining information from multiple sources, not only sensor data, but also information stored on the Internet, can lead to a better understanding of the problem to be solved, such as healthcare or vehicle mobility. For example, traffic in cities is partially dependent on local holidays. These multi-source data have resulted in the growth of some research fields, such as information fusion, intelligent environments, and ubiquitous computing. The insights from analyzing these multisource datasets can be applied to real-world problems such as tourism management, economics, and financial information systems [28].

The number of studies with smart city data has grown exponentially in recent years. The most important cities have deployed sensor networks and IoT platforms. The data obtained by these sensors have led to numerous studies in several areas, such as traffic behavior [43, 37, 52, 46]. However, most solutions that try to cluster different traffic behavior do not have additional information, such as the residence of vehicle owners, to provide additional insight into the explainability of the clusters. Furthermore, this smart city trend has yet to reach small villages, and the solutions found for large cities do not always apply directly to small villages. For example, solutions monitoring traffic behavior in large cities with numerous streets and several traffic lines in some avenues do not extrapolate to villages with 6 or 7 mostly pedestrian streets and only one road with one line in each direction. Additionally, even if we try to add some explanation to the behavioral cluster in smart villages, the residency of vehicle owners is not straightforward. Due to the recent movement on moving from cities to villages and retrying or spending long periods on second residences, the actual residence information is fuzzy in rural villages.

This work proposes a clustering based on vehicle behavior in small villages, with information from license plate recognition (LPR) devices and owners' residences, among others. We applied the study directly to each individual (vehicles) and defined their spatio-temporal behavior based on their spatial frequencies of visitation. To that end, we fused several datasets and calculated new valuable variables such

[1] https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/

as the time spent in the area; total distance traveled there, etc. We study the popularly used clustering algorithms to draw conclusions on which of them performs better on the problem under consideration. In particular, we used a pipeline to analyze the particularities of the data through several visualization tools, and we explained, based on the data, the optimal normalization and clustering algorithm that best groups the different behaviors of the vehicles. We will focus on the importance of selecting the optimal normalization algorithm and its influence on the results. Additionally, we analyzed the results with residential information and determined the variables that most influence each cluster. With this information, we explained the behavior pattern of each cluster. Our pipeline comprises eight steps: data collection, cleaning, fusion, normalization, dimensionality reduction, clustering, evaluation, and visualization. Finally, we applied the proposed pipeline to a touristic rural region, with the problems mentioned above of a single small road and the lack of reliable residency information.

The findings of our research hold significant implications for policymakers, particularly in managing tourism flows in smart cities and villages. By integrating multiple data sources and employing appropriate clustering and normalization algorithms, our proposed pipeline enables a comprehensive analysis of mobility patterns. The enriched dataset, with its valuable insights into different patterns, can inform policymakers about the dynamics of tourism flows in specific areas. This knowledge is crucial for managing and optimizing transportation systems, infrastructure development, and resource allocation to accommodate the needs of both residents and visitors. By understanding the traffic patterns among residents and tourists in a rural touristic area, policymakers can devise targeted strategies to enhance visitor experiences, reduce congestion, and mitigate the environmental impact associated with tourism activities. The data analysts and policymakers can utilize our research to select suitable algorithms and gain a deeper understanding of mobility patterns, thus facilitating evidence-based decision-making and the effective management of tourism flows in environmentally sensitive areas.

1.1 Motivation

In conducting a literature review, I observed a significant gap in the analysis of smart villages and car behavior within these environments, compared to the abundance of articles focused on smart cities. In addition, there was limited exploration of normalization techniques in this context. This realization sparked a personal interest and desire to tackle the challenge of working with sensors and creating a comprehensive database from a simple one constructed from timestamps and license plate data.

Given my background, I was able to combine different fields of expertise, combining my previous knowledge in computer science with those acquired during my master's degree, such as business intelligence, data visualization and information extraction for policymakers. During my degree, I did a simple internship on clustering analysis, and dealing with the topic of unsupervised learning in the master's degree, aroused even more my interest to go deeper into this topic.

As an additional motivation, the completion of this Master's Thesis (MT) is also linked to Smart Poqueira[2] a subproject included in the project "Thematic Center on

[2] https://wpd.ugr.es/~smartpoqueira/

Mountain Ecosystem & Remote sensing, Deep learning-AI e-Services" (LifeWatch-2019-10-UGR-01), linked to the analysis of different aspects related to the conservation of the Sierra Nevada National Park through advanced digital systems. The project has been co-funded by the Ministry of Science and Innovation through the ERDF funds of the Pluriregional Operational Program of Spain 2014-2020 (POPE), LifeWatch-ERIC action line, with the co-funding of the Provincial Council of Granada and the University of Granada. The heads of the Provincial Council and the University of Granada have pointed out that this commitment to the use of digital technology and innovation will make it possible to efficiently manage numerous data on the behavior of its visitors and, thus, implement solutions that improve the quality and sustainability of the visits received.

In addition to the connection with the Smart Poqueira Project, this MT is closely related to several subjects learned in the Master's program, providing a solid foundation and relevant knowledge for its development. The following are some of the areas from the master's degree courses that have a strong relationship with this dissertation:

- **Analysis and Inference in Business Processes**: This course has provided me with a background in machine learning techniques and validation metrics. By leveraging this knowledge, I have been able to develop intelligent systems that extract relevant information and patterns for decision-making in the field of sustainable tourism.

- **Business Intelligence**: Data analysis related to business strategy is a key element in implementing sustainable tourism solutions. In this subject, I have improved my skills in collecting, analyzing, and visualizing data, enabling me to make well-informed decisions for tourism management in the region of Alpujarra.

- **Data Bases for Business Processes**: Effective management of the data collected is crucial to the success of the project. By studying this subject, I have acquired the knowledge necessary to design and ensure the integrity and usefulness of the information collected.

- **Big Data and Sustainability**: My understanding of big data and its implications for sustainability in the tourism industry was fostered through the subject of **Introduction to Management and Technologies in Business Processes**. Efficient management of collected data, its analysis, and the generation of insights from it will be key elements in making informed decisions and developing strategies that foster sustainability in the tourism sector.

- **Project Management and Planing**: Effective communication and planning are essential for successful project execution. In this course, I have learned how to establish clear lines of communication with suppliers and stakeholders. In addition, this subject has given me project management techniques that support the design of project requirements and timelines essential in project development.

- **Business Strategy and Internationalization in Technologically Advanced Environments**: In this course, I have learned that, today, organizations and businesses face a rapidly changing and highly competitive global environment.

> Policymakers and business owners must be agile, flexible and be able to obtain the necessary information for strategic decision-making, which facilitates higher levels of current and future competitiveness. The integration of technologies into business processes facilitates responsiveness, cost reduction, improved productivity and enhanced services.

These subjects provide a solid foundation of knowledge and skills that will be directly applied in the development of the MT. Additionally, other competencies from subjects in the program have been developed through teamwork, which has made this work possible. **Collaborative Systems and Workflow Management** (communication tools with teams and suppliers), **Dashboards and Multidimensional Systems** (designing KPIs and dashboards linked to the project), and **Design and Access to Web-Based Information and Content Management** (development of the project's website) have all contributed to the completion of this work. The combination of the connection with the Smart Poqueira project and the relevance to key subjects in the Master's program ensures the significance and applicability of this work, making a significant contribution to the advancement and development of sustainable tourism in Alpujarra.

1.2 Objectives

The objective of this MT is to investigate and analyze traffic behavior in smart villages using license plate recognition (LPR) devices and considering the residence of the owners. The specific objectives are as follows:

- Study clustering algorithms, machine learning pipelines and normalization algorithms; and the suitability of using specific algorithms for specific data distributions.
- Review previous works on clustering applied to traffic behavior with LPR.
- Design and deploy an infrastructure to collect data on vehicle movements in smart villages.
- Examine, in our use case, the correlation between vehicle behavior and provenance, visitation frequency, length of stay, and seasonality.
- Evaluate and compare different normalization techniques for preprocessing the collected data.
- Explore various clustering algorithms and techniques and select the most appropriate ones based on the visualization and explainability of the results.
- Evaluate the performance of the proposed clustering pipeline and selected clustering and normalization algorithms.

By achieving these objectives, this work aims to contribute to the understanding of traffic behavior in smart villages and provide valuable information for the design of effective traffic and tourism management systems and policies in the area.

1.3 Thesis Outline

This chapter has presented the challenges that face the study of traffic in smart villages and the motivation of this dissertation. In Chapter 2 related work is summarized. Chapter 3 presents the theoretical bases discussed throughout the MT, describing the main normalization and clustering algorithms and metrics, as well as the demographic context of the area where the study was conducted. Chapter 4 presents the unsupervised learning pipeline, including the sensor's setup and the different sources of information used to construct the dataset. Chapters 5 show the analysis of the results. Finally, Chapter 6 concludes the work and Chapter 7 presents the limitations and future work related to the project.

Chapter 2

Related Work

2.1 Information Fusion in Smart Cities for Touristic applications

The concept of information fusion has been applied to the specific problem of tourism flows and smart cities. These approaches use advanced data analysis techniques to combine multiple sources of information, providing valuable insights for developing smart tourism applications in cities and designing sustainable environments. Smart city applications are built on top of data, and data fusion has provided a wide variety of techniques to improve the input data for an application [35]. Examples of these techniques include data association, state estimation, unsupervised machine learning, or statistical inference. For example, combining different tourist information has been used to predict the national tourist flow in Spain with graph neural networks [60]. The data used in the solution are composed of tourist infrastructure information, such as camping and tourist housing from the data sources OpenStreetMap and the National Institute of Statistics of Spain (INE); reports released by the Spanish Ministry of Transportation (SGTM); and human mobility data including the number of movements between administrative areas per hour extracted from geotagged Twitter data. Most of these applications are focused either on user recommendations or tourist flow, but little attention has been paid to studying the individual behavior of the tourist inside an area (for a detailed survey, see [18, 35]).

2.2 Clustering and mobility patterns

The increasing deployment of IoT platforms in smart cities has boosted the proliferation of sensors, including those that monitor traffic. These sensory data allow us to analyze vehicle behavior. The most common works in this area are to analyze mobility patterns in order to improve traffic congestion [43, 52], and to aggregate vehicles to obtain useful conclusions for urban management [12, 37].

To infer mobility patterns from raw data, unsupervised ML is widely adopted. In particular, various industries use clustering algorithms to categorize data into distinct groups based on similarities, differences, and patterns without prior knowledge. Clustering analysis is used to detect behavioral patterns in the field of pedestrian-vehicle mobility, and in the field of indoor-outdoor (IO) positioning systems [39]. Some works use partitional clustering to analyze data. For example, in [75], the ISODATA clustering algorithm is used to cluster mobility patterns and a decision tree is used to create decision rules between the attributes and the labeling obtained from

the clustering. In [65], they propose a K-Means clustering framework combined with other processes such as dimensionality reduction and feature extraction to classify tourists and locals based on the data generated by each individual's mobile phone signaling data and their movement through the area. Hierarchical clustering [49] has also been used to segment time series related to vehicle mobility, with the objective of predicting areas where there is a higher risk of accidents. Other commonly used clustering algorithms are density-based, as they can be adapted to problems where irregular behavior occurs within the population. In [7], the authors propose an alternative version of the DBSCAN clustering algorithm to detect collective anomalous human behavior from large amounts of pedestrian data in smart cities. The algorithm uses an iterative search process and aggregation of achievable density data points to form clusters, culminating in a global approach to identify behaviors of particular interest in the population. Density clustering techniques also allow the analysis of movement in areas where some specific transport behaviors are known but where more information about particular groups is desired. In [4], a modified version of DBSCAN (iterative and multi-attribute) was used to cluster the different areas of the port, with the aim of improving organization and resolving port congestion. Algorithms such as GaussianMixture are used to perform segment analysis, where individuals are defined by their movement routines, and the data is related to the frequency and period of stay in different areas. From the movement information provided by smart cards, several papers apply this algorithm to identify market segments based on temporal travel patterns [14], define tourist patterns based on frequency and areas where transactions are made [26] or identify changes in functional areas of cities over time [68].

Few works related to clustering analysis in mobility use LPR cameras as the main source of information [73]. For example, [73] analyze the commuting patterns constructing the spatio-temporal similarity matrix using the dynamic time warping (DTW) algorithm; and afterward, analyze the characteristics of commuting patterns with the density-based spatial clustering of applications with noise (DBSCAN) algorithm. Similarly, [74] analyzes the change in traffic patterns during the pandemic using K-Means. However, none of these works combine LPR data with vehicle provenance nor study the touristic behavior of the vehicle. Likewise, none of them compares the suitability of using different clustering algorithms.

2.3 Traffic management and business strategy in villages

Traffic management and business strategy in rural villages is a crucial topic for the development of rural policies in Europe. The implementation of smart villages presents distinct challenges in both central and peripheral rural areas at the European and national levels. It is necessary to have an integrated vision that takes into account the specific issues, needs and expectations of each country. In this context, Spain finds itself in a situation of dual rural periphery: European and national. There are two paths towards smart villages: the horizontal connection between territories and the businesses that promote diversification [47]. Depopulated areas are attractive for leisure, work, and retirement. The challenge lies in developing intelligent and competitive policies that encompass all European rural areas, as well as implementing intelligent and competitive policies and strategies in depopulated zones. These policies may include the development of sustainable indicators to avoid overtourism [51], while businesses can leverage data on the frequency and duration of

tourist visits to determine optimal opening hours and the types of goods and services to offer [71]. Data on the modes of transportation used by visitors can also provide insights into the necessary transportation infrastructure to support businesses in the area.

Analyzing traffic patterns can also help businesses to better understand the impact of tourism on their operations. By identifying peak tourist seasons and the types of tourists that frequent the area, businesses can adapt their strategies accordingly [11]. For instance, they can adjust their marketing campaigns to target specific groups of tourists or offer promotions during off-peak seasons to attract more visitors. The highlights the importance of data-driven decision-making in developing strategies that are tailored to the needs of visitors and the local community [45]. Such analyses also provide policymakers with insights to understand mobility patterns in environmentally sensitive areas, ultimately leading to better planning and management of transportation infrastructure.

Chapter 3

Fundamentals

3.1 Main clustering algorithms

Unsupervised machine learning automates the knowledge discovery process without the need for labeled data or previously classified data [35]. Techniques such as clustering and anomaly detection fall under this category. Clustering is an unsupervised ML technique that aims to find patterns in observations of events. Most taxonomies group the algorithms into at least five categories [27], although we have identified seven, as some of them did not fit in the 5 elements taxonomy:

Partitional Clustering: This clustering technique decomposes a dataset into distinct clusters through an iterative process of distance calculations between individuals, and typically uses centroids. Examples of algorithms that utilize this technique include K-Means and MiniBatchKMeans, which is a scalable version of K-Means that updates clusters using small random batches until convergence is achieved [6]. Another algorithm that falls into this category is ISODATA [42], which employs iterative self-organizing data analysis.

Hierarchical Clustering: This clustering method constructs clusters in either an agglomerative or divisive manner by adding or removing individuals, respectively. BIRCH [77], an algorithm that uses an unbalanced height tree to dynamically split data points, is a popular example of hierarchical clustering.

Density-based Clustering: This technique identifies dense regions of objects in the data space separated by low-density regions. It is known to handle noise well and adapt to arbitrary shapes in the data. The algorithm most commonly in this category is DBSCAN [20], along with improved versions such as OPTICS [1] and HDBSCAN [41], which compute a density function for each cluster found. Other examples include Mean-shift [16], which creates clusters based on regions of maximum density attraction and can be considered a version of K-Means using density functions, making it adaptable to arbitrary shapes of clusters.

Distribution-based Clustering: This technique creates clusters based on the probability that each individual belongs to the same distribution, the Gaussian distribution is the most widely used distribution based on the expectation maximization algorithm [72]. These algorithms result in Gaussian Mixture models, which are also classification algorithms. In some cases, they are a generalization of K-Means, with each individual having a probability of belonging to each cluster.

Clustering Category	Algorithms	Application	Related work
Partitional	K-Means, MiniBatchKMeans, ISODATA	Target classes, analyze patterns	[74, 65, 75]
Hierarchical	Agglomerative clustering, Divisive clustering, BIRCH	Behavioral patterns, feature extraction	[49, 76, 34]
Density-based	DBSCAN, OPTICS, HDBSCAN, MeanShift	Complexity reduction, anomaly detection	[52, 4, 73, 7]
Distribution-based	Gaussian Mixture	Density estimation, outlier detection	[14, 26, 68]
Grid-based	STING, WaveCluster, CLIQUE	Spatial-based segmentation	Not found
Message passing-based	Affinity Propagation, IWC-KAP, ScaleAP	Clustering indoor location patterns	[40, 78, 44]
Spectral	Spectral Clustering, ASC	Graph partitioning, image segmentation	[55, 48, 36]

TABLE 3.1: Examples of works using clustering to infer mobility pattern in 2020-2023.

Grid-based Clustering: This clustering approach involves dividing the space into a finite number of cells, followed by defining clustering operations within the quantized space. Some popular algorithms that utilize this method include STING [67], WaveCluster [62], and CLIQUE [21].

Message-Passing Clustering: This category of clustering creates clusters by exchanging messages between different data points until convergence. An example of this approach is the Affinity Propagation (AP) algorithm [22], which has been further improved by proposals such as IWC-KAP [61] and ScaleAP [63].

Spectral Clustering: This method uses the spectral radius of a similarity matrix of the data in a multidimensional problem. Dimensionality reduction techniques, such as Principal Component Analysis (PCA), are used to obtain a linearly separable problem. There are different versions of Spectral Clustering algorithms, depending on how the eigenvectors are selected from the Laplacian of the similarity matrix [66]. Newer versions, such as Attributed Spectral Clustering (ASC), improve the degree of affinity between nodes in the same density region [8].

Table 3.1 shows the main algorithms in each category described in this section, and examples of applications for each algorithm, in the field of mobility pattern analysis in the last 3 years (2020-2023).

3.2 Clustering performance

Clustering is difficult to evaluate, as we do not know the ground-true, i.e. we do not have labeled data, to evaluate whether the clustering algorithm has grouped each individual in the right cluster. However, there are some metrics that could give some insight into how good the clustering is based on the distances between groups or the balance between groups, or the density of individuals in each group. The three most popular internal evaluation metrics in the literature [38] are silhouette coefficient, calinski-harabasz score, and davies-bouldin index. All of these metrics are based on distances between data points and are commonly used to evaluate the effectiveness of virtually any clustering algorithm, working especially well in algorithms

that work with distances, such as those included in the hierarchical, partitional, or spectral categories.

- **Silhouette Coefficient** (SC): measures the similarity, based on distances, of an individual to its own cluster compared to other clusters [59]. The coefficient value ranges between $[-1, 1]$, where 1 represents a good clustering division and a value close to -1 represents a poor division.

  The silhouette coefficient of one data point $i \in C_i$ is:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \; if \; |C_i| > 1, \;\; s(i) = 0 \; if \; |C_i| = 1 \tag{3.1}$$

  Where C_i represents the cluster to which the data point i belongs, and $|C_i|$ is the cluster size, i.e. the total number of points contained in C_i.

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} ||j - i||, \;\; b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} ||j - i|| \tag{3.2}$$

  Where $a(i)$ is the average distance between a data point i and all other data points in the same cluster C_i, and $b(i)$ is the average distance between a data point i and all data points in the nearest cluster other than C_i.

  For n the total number of data points, the global silhouette coefficient is defined as:

$$SC = \frac{1}{n} \sum_{i=1}^{n} s(i) \tag{3.3}$$

- **Calinski-Harabasz Score** (CH): like the silhouette coefficient, measures how similar an individual is to its group relative to other groups [13]. A higher value minimizes the intracluster covariance of individuals and maximizes the intercluster covariance. In cluster analysis, the within-group variance and between-group variance can be calculated by sum-of-squares within a cluster (SSW) and sum-of-squares between clusters (SSB) respectively.

  Sum of Squared Within (SSW): minimizes the distance between individuals in the same cluster (cohesion).

$$SSW = \sum_{i=1}^{k} \sum_{i \in C_i} ||i - m_i||^2 \tag{3.4}$$

  where k is the number of clusters, i is a point of cluster C_i and m is the centroid of a cluster C_i.

  Sum of Squared Between (SSB): maximizes the distance between individuals from different clusters (separation).

$$SSB = \sum_{j=1}^{k} |C_j| ||m_j - \bar{x}||^2 \tag{3.5}$$

where k is the number of clusters, $|C_j|$ is the number of elements in a cluster j, m_j is the centroid of the cluster j and $\bar{x}$ is the mean of the dataset.

The CH score is the division between both variances:

$$CH = \frac{SSB(n-k)}{SSW(k-1)} \tag{3.6}$$

where k is the number of clusters and n is the sample size.

- **Davies-Bouldin Index** (DBI): Small values indicate compact clusters with well-differentiated centers that are far apart from each other. [17].

$$DBI = \frac{1}{k} \sum_{i=1, i \neq j}^{k} \max\left(\frac{\sigma_i + \sigma_j}{||C_j - C_i||}\right) \tag{3.7}$$

where k is the number of clusters, σ_p is the average distance between each point in a cluster p and the centroid of its cluster (with $p \in \{i, j\}$) and $||C_j - C_i||$ is the distance between the centroids of the two clusters.

The distance-based metrics discussed above may not be suitable for algorithms that use the Expectation Maximization (EM) method, such as the GaussianMixture algorithm. This is because the EM method models the data using probability distributions rather than distances between data points. Therefore, we might get some imprecision when comparing the performance of algorithms of this type if we use these metrics. Instead of using distance-based metrics, distribution-based algorithms typically use statistical criteria to determine the optimal number of clusters or components that best fit the data. These metrics aim to balance the trade-off between an ideal clustering and the number of parameters used in the model, with a penalty for models that have too many parameters. One of the scores used is the information criterion (IC), a penalized likelihood function that includes a negative log-likelihood function and an aggregate penalty term that increases with the number of parameters in the model.

$$IC(K) = -2 \cdot l(\hat{\Psi}|C) + d(K) \cdot a_n \tag{3.8}$$

where $\hat{\Psi}$ is the estimate of the parameters of the K-component mixture model, $d(K)$ the number of parameters of the K-component mixtures model, $l(\hat{\Psi}|C)$ the log-likelihood function, n the sample size, and a_n an increasing function. The optimal number of clusters is the one that minimizes the IC.

The following are two of the best-known variations of information criteria used in the literature [30]:

- **Akaike information criterion** (AIC): AIC is a particular specification of the general information criterion (IC), in which $a_n = 2$. This criterion is known to

overestimate the order of the model.

$$AIC(K) = -2 \cdot l(\hat{\Psi}|C) + 2 \cdot d(K) \tag{3.9}$$

- **Bayesian information criterion** (BIC): Tries to overcome the overestimate of AIC. The penalty term depends on the sample size n, so as $n \to \infty$ the penalty is larger and does not overestimate the order of the mixture as much as AIC does [5].

$$BIC(K) = -2 \cdot l(\hat{\Psi}|C) + \log n \cdot d(K) \tag{3.10}$$

3.3 Principal Component Analysis

The Principal Component Analysis (PCA) method reduces the dimensionality of a dataset in order to simplify the complexity of the ML analysis with an elevated number of variables. This method condenses the information provided by multiple variables $(X_1, \ldots, X_p)$ from a given sample into a smaller number of variables, finding a number s of underlying factors that explain approximately the same variance as the original variables with $s < p$. Each of the new variables $(Z_1, \ldots, Z_p)$ are called principal components, which correspond to an eigenvector of the covariance matrix associated with the data. These new variables are linear combinations of the original variables.

Each principal component (Z_i) is defined as a normalized linear combination of the original variables (X_i) under a variance maximization problem, indicating that the new component best summarizes the information contained in the original variables. We define each Z_i as:

$$Z_i = \Phi_{1i}X_1 + \Phi_{2i}X_2 + \ldots + \Phi_{pi}X_p \tag{3.11}$$

Each Φ represents the weight or importance that each variable X_i has in each Z_i and, explains the information collected by each of the principal components. To compute the first principal component of a dataset with n observations and p variables, first, centralize the variables so that they have zero mean. Then an optimization problem is solved to find the values of Φ that maximize the variance, using the eigenvectors of the covariance matrix. Once the first component is obtained, the second component is calculated following the same process, but adding the condition of no correlation with the already calculated components [31].

It is advisable to apply prior normalization to the data, since this method is highly sensitive to variables of different scales. Furthermore, the PCA only works with numerical data, so it is necessary to perform a previous preprocessing on categorical variables that may exist in the input dataset [57].

3.4 Normalization

The existence of attributes at different scales and measured in different units increases the influence of some variables over others in the clustering process. Normalization compresses or expands the values of each variable to fit them in the same range of values, normally [0,1], or [-1, 1], making them comparable in subsequent processes (PCA or ML algorithms). The choice of the normalization algorithm usually depends on the specific application and the dataset used, as different methods may yield different results and interpretations. For example, in clustering analysis, normalization can be particularly important for comparing similarities between characteristics based on certain distance measures. Among the most commonly used normalization methods in the literature are min-max normalization and Z-score standardization [29, 50]. In the context of PCA, z-score standardization is often preferred over min-max normalization. The z-score standardization method handles outliers better because it uses the standard deviation to scale the data, rather than a fixed range as in min-max normalization, where outliers can significantly affect the overall scale of the dataset. However, it is important to note that being more sensitive to outliers does not always work well for all datasets. We have also tested two other methods that are commonly used in the literature [54, 3] and occasionally produce better results than the two described above.

1. **Min-max normalization:** Uses the minimum and maximum in the attribute domain to normalize the values to the interval, $[0, 1]$ keeping the distances for each data point X.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3.12}$$

2. **Z-score standardization:** scales the values so that the mean (μ) of the data domain is 0 and the standard deviation (σ) is equal to 1.

$$X' = \frac{X - \mu}{\sigma} \tag{3.13}$$

3. **Median Absolute Deviation (MAD) normalization:** normalizes the data such that the median of each attribute is 0 and the median absolute deviation is equal to 1. The formula for MAD normalization is shown below:

$$X' = \frac{X - median(X)}{MAD(X)} \tag{3.14}$$

   Where $median(X)$ is the median of the values in attribute X, and $MAD(X)$ is the median absolute deviation of X.

4. ℓ^2 **normalization:** normalizes the data by dividing it by its Euclidean norm. This ensures that all feature vectors have the same length and is commonly used in machine learning and information retrieval. The formula for ℓ^2 normalization is shown below:

$$X' = \frac{X}{||X||_2} \tag{3.15}$$

Where $||X||_2$ is the Euclidean norm of, X given by $\sqrt{\sum_{i=1}^{n} X_i^2}$.

3.5 Dataset geometry

In mathematics, a Riemannian manifold is a geometric object that can be described locally as Euclidean space. Curvature is an intrinsic measure of a manifold, indicating how much the manifold curves at each point. In this context, we say that a manifold is flat if its curvature is zero at all points, that is, if the manifold is locally indistinguishable from a flat Euclidean space. If the curvature is not zero at any point of the manifold, the manifold is non-flat. In data analysis, we refer to flat and non-flat geometry as the measurement of distances between points by Euclidean or non-Euclidean geometric methods, respectively. In flat geometry, the distance is measured following a straight line between two points, while in non-flat geometry, the distance is measured following a curve. We can detect whether our data follow flat or non-flat geometry by representing the data in a scatter plot, where each point represents an individual in the population. Visually we can only represent 3 dimensions, which normally are the most representative variables of the cluster, or the firsts principal components of a dimensional reduction algorithm. If the resulting figure shows a roughly circular, rectangular, or elliptical shape, the data are likely to follow a flat geometry. However, if the figure has an irregular, twisted, or folded shape, the data are likely to follow a non-flat geometry. From different studies [24], it has been found that partitional or distribution category clustering algorithms work best with data cases that follow a flat geometry, while density-based and message-passing algorithms work best with non-flat geometries.[1]

3.6 Case study: smart villages

Recent years have seen a growing trend of urban exodus, with many people leaving the cities in search of a quieter life. This development is largely due to the positive perception of the quality of life in rural areas, which offer a number of amenities attractive to those seeking a more relaxed lifestyle. In addition, with the advent of COVID-19, there has been a significant increase in the urban exodus, as many people have opted to live in less populated environments [69]. With the rise of telecommuting, this trend is likely to continue in the future, with an increase in the number of people choosing to live in villages while working from home for companies in large cities. These migratory flows include both foreign immigrants and the arrival of resident citizens from other parts of the country, attracted by new conditions such as living in a cheaper and less crowded [53] environment. Also, noteworthy are the groups of retirees (both foreigners and nationals), who move to the countryside to acquire more comfortable homes, leading a quiet lifestyle that allows them to enjoy higher levels of social and environmental capital [70, 58]. Most of these newcomers to the rural areas do not register their vehicles in their new residences.

[1] https://scikit-learn.org/stable/modules/clustering.html

The Alpujarra Granadina is a region located in the Sierra Nevada National Park in Granada, Spain. This region is made up of 32 municipalities with an average of fewer than 1000 inhabitants that enjoy a great tourist attraction with visitors of different nationalities [19]. The Alpujarra is an area that attracts foreign and national retirees, from other regions further north, who spend several months of the year there, avoiding the colder winter months. There are also new groups of people (neo-rural) who, motivated by environmental movements or simply the search for a quieter life, come from other parts of the country or other countries to experience an exotic village and become residents of the area for several months [10]. Both groups of individuals; retirees and neo-rurals, face a transition or permanent period, leaving the address of the vehicle registered to their previous residence. Following the concept discussed in [58], we will call these groups of "false residents" non-registered residents, since they do not have a residence permit but do have a dwelling or habitable accommodation during the long period of stay. Therefore, within the group of residents, we will distinguish between those who are registered in the study area (registered residents) and those who, despite behaving as residents and having their own homes, are not registered in any of the municipalities (non-registered residents), but who represent an important part of the population of the Alpujarra. Located on the southern slope of Sierra Nevada and within the northern part of the Alpujarra, is the Barraco de Poqueira, a region formed by the municipalities of Pampaneira, Bubión and Capileira [19], and within which our case study is located. Preserving the ecosystem of this region is essential because it is situated in a natural park near a national park with unique biodiversity. Hence, we selected this region because it is essential to balance the wealth that brings tourists to the zone with the pollution generated by vehicles. Understanding the patterns of the vehicles in the zone is the first step to generating suitable policies to preserve the area's sustainability.

3.6.1 Addressing challenges in small tourist villages

Small tourist villages pose a significant challenge to policymakers due to the presence of non-registered residents. These non-registered residents, who are not accounted for in official statistics, can have a significant impact on the local economy and the provision of public services. In many cases, these non-registered residents are individuals who own or rent properties in the village but are not officially registered with local authorities as permanent residents [32]. Policymakers may need to explore creative solutions, such as offering tax incentives or social programs, to encourage non-registered residents to participate more actively in the local community. Another challenge for policymakers is finding ways to promote sustainable tourism in the area, by creating personalized experiences based on different seasons or the origin of the tourist flow [64]. Consequently, policymakers are striving to identify their needs and demands and provide them with appropriate services.

One way to address these challenges is to use data-driven methods to better understand the characteristics and behaviors of individuals who visit the village [33]. Data-driven methods have a huge potential to comprehend visitor behavior and, based on that, develop policies and services that meet their needs. For instance, the use of mobile device data or any other information captured by sensors can provide insights into the movements and activities of individuals in the village [26]. Machine learning algorithms can then be employed to analyze this data and segment behaviors into different clusters. Policymakers can leverage this information

to develop more effective policies and services that cater to common needs and enhance the overall economic and social well-being of the village. By addressing these challenges, policymakers can ensure that small tourist villages remain sustainable communities for both visitors and residents.

3.7 Smart Village Platform Design and Deployment

The main objective of the Smart Poqueira project is the development of sustainable tourism in Alpujarra. The first step to develop a sustainble tourism is to study the current tourism. To that end we use sensorization tools to gather information on vehicles, provenance, and other relevant aspects, as well as the analysis of related data.

In addition to the construction of a ML pipeline for data analysis, it is necessary to highlight my role in shaping the project's design and my contributions to various essential pre-tasks presented in this work. These include the decision-making process regarding camera placement, project management support, communication with LPR suppliers, and analysis of advantages and decisions made.

Camera Placement

As a member of the project's design team, I participated in the task of identifying and determining optimal locations for camera installation. To accomplish this, we discussed various options and analyzed advantages and alternatives that would minimize costs while maximizing the information extracted by the cameras (discussed in Section 4.1). Furthermore, we monitored the proper functioning of the cameras in the following months after installation, considering factors such as viewing angles, lighting conditions, and reporting incidents.

Communication with Suppliers

As part of the project, a collaboration was established with the suppliers of the implemented LPR cameras. I served as one of the primary liaisons between our team and the external suppliers. My role involved establishing clear and effective communication, and ensuring that project requirements and expectations were adequately conveyed. I coordinated meetings with suppliers' team leaders to discuss progress, address issues or concerns, and ensure the promised quality was met.

In summary, my involvement in the university project encompassed the configuration of the design, camera placement, project management support, and communication with external suppliers. Through my contributions in these areas, I aimed to ensure efficiency, effectiveness, and overall project success.

Chapter 4

Clustering Pipeline

To analyze vehicle behavior, we have designed an information fusion pipeline, which divides the analysis into different stages. Each stage fulfills its own objective of data processing and extraction of information that will be relevant in the subsequent stage. In general, the pipeline begins with the extraction and collection of data from heterogeneous sources and finally produces a grouping result from a clustering model based on the decisions we make along the pipeline (see in Figure 4.1). The technologies used to implement and execute the different stages of the pipeline come from the fields of ML and data analysis. In Table 4.1, we can visualize a description of the different stages proposed in the pipeline and the experimental values considered in each of them. The pipeline consists of the following stages: data collection, data cleaning, data fusion, preprocessing, dimension reduction, clustering, evaluation, and visualization.

4.1 Data collection

The data collection phase handles the collection and storage of our different data sources. This process involves collecting data from different sensors and other data sources, such as Web pages or databases. In our case, we collected the data from LPR cameras and from specific databases we collected: vehicle information, demographic and economic, national calendar, and geographic data.

Regarding vehicle tracking infrastructure (LPR cameras), data is collected by four devices equipped with vehicle detection sensors. These devices are Hikvision LPR IP cameras with Automatic number-plate recognition (ANPR) based on Deep Learning. The devices have a 2MP resolution, 2.8-12 mm varifocal optics, and IR LEDs with a range of 50 m.

To cover the entrances and exits of each village in the target area, we strategically positioned the four cameras, as shown in Figure 4.2. The locations are (i) entrance to Pampaneira from the western part of the Alpujarra, (ii) entrance to Pampaneira from the eastern part of the Alpujarra, (iii) entrance to Bubión via a single road, and (iv) entrance to Capileira via a single road. By taking advantage of the road structure, we can monitor the mobility of all vehicles that circulate in the Poqueira area using only four LPRs.

The main objective of these cameras is to track vehicles entering and leaving each of the villages, providing detailed knowledge of mobility in the Poqueira area. Using only four cameras also helped minimize the cost and complexity of the system while

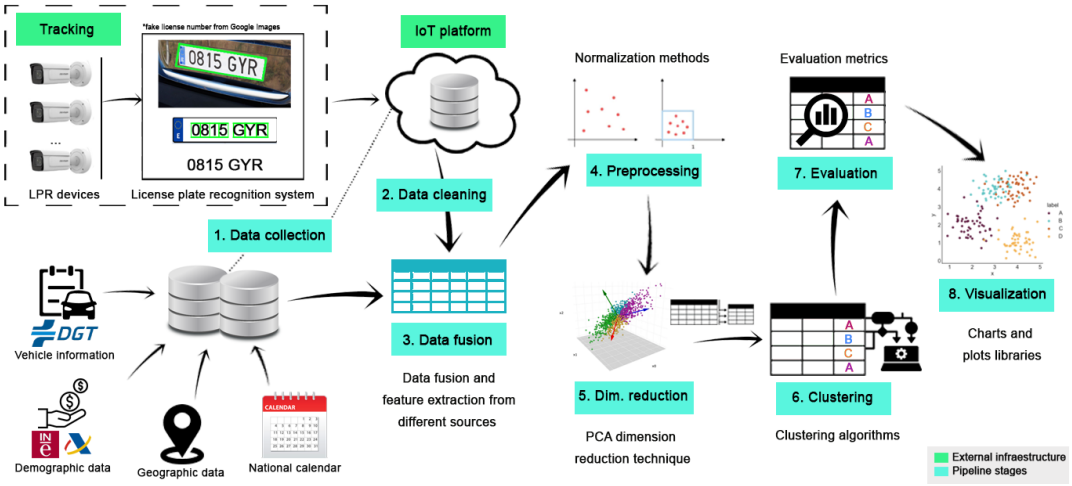


FIGURE 4.1: Overview of the clustering pipeline.

Stage	Configuration parameters	Experimental values
Data Collection	Data collection from different sources	Storage in own DB and external IoT platform
Data Cleaning	Recovery and treatment of lost data	1. License plate matching 2. Recover movement of vehicles not detected by any camera in their total route
Data Fusion	Fusion of information data and feature extraction	Detailed process in Table 4.2
Preprocessing	Normalization methods	Min-max normalization, z-score standarization, MAD normalization, ℓ^2 normalizacion
Dimension reduction	Dimension reduction techniques	Principal Component Analysis (PCA)
Clustering	Clustering algorithms	K-Means, MiniBatchKMeans, Agglomerative clustering, BIRCH, DBSCAN, HDBSCAN, MeanShift, Gausian Mixture, Spectral Clustering
Evaluation	Evaluation metrics	Silhouette, Davies–Bouldin, Calinski–Harabasz, number of clusters, Bayesian Information Criterion, Akaike Information Criterion
Visualization	Visualization plots	box plot, scatter-plot, elbow method, PCA variance plot

TABLE 4.1: Configuration of each stage of the pipeline with the values used in this study.

FIGURE 4.2: Setup of the 4 LPR that obtain the data from the license plates of the vehicles.

still capturing the necessary data. The information collected by the cameras is stored on a cloud platform.

The rest of the data were collected from different databases. Vehicle information from a private dataset of the National Traffic Department. The demographic and economic information from the Internet website of the National Statistics Institute and the Tax Agency. the national calendar and the geographic data are from Python libraries.

4.2 Data Cleaning

In the field of the IoT, the production of sensor data can often be inaccurate and lead to the loss of some records. In our case, we present two cleaning steps for the main dataset (LPR cameras). The first step, “license plate matching", aims to reduce the error rate of incomplete or wrongly detected license plates by the LPR devices to maintain consistency between vehicles with the same license plate. About 2% of the stored 1,050,760 records have missing values in the license plate number. For example, if we have a record with a correct license plate 0000AAA, and another record with the value 0#00AAA, missing the second digit, we could, by probability, infer that both records belong to the same plate number and assign the correct value, 0000AAA, to both records. In our case, we assign the same plate number to all those records whose license plate matches at least four characters out of seven in the same position. The second step, “route recovery", aims to reduce the percentage of vehicles not detected by any LPR device. These errors occur when the camera does not detect a vehicle that passes through the road. This error is difficult to detect, but in our setup, if a vehicle moves on the road from camera 1 to 3, and camera 2 (in the middle of the unique road connecting cameras 1 and 3), does not detect the car, we could infer that the car has passed through camera 2. In our process, if the vehicle is detected in less than 30 minutes in two non-consecutive cameras, our system infers

that the vehicle is still in the area and calculates its time of stay based on the new registered values.

4.3 Data Fusion

The area where we conducted the experiments, as presented in Section 3.6, records individuals with diverse behavioral patterns and experiences a large flow of tourists from different parts of Spain and elsewhere, each with unique mobility characteristics that vary by region of provenance. Combining data from provenance, mobility in the area, and the holiday calendar offers the opportunity to gain an understanding of the region, its inhabitants, and visitors. This section explains each source of information and the feature extraction and construction process of each dataset to allow the merging. We will detail the structure and variables obtained for each data source, creating a joint database. Table 4.2 schematically shows the information fusion process we have followed.

License Plate Recognition Data

The LPR cameras described in Section 4.1 return information on four variables: the vehicle license plate (license_plate), the time stamp (time_stamp), and the direction (type) for each camera defined by an identifier (camera_id). The dataset contains information for nine months (February to October 2022). On these data, we perform the data cleaning that we defined in Section 4.2. We change the original license plate value to an integer value that functions as the vehicle identifier. In total, we have 1,050,760 records, of which 25.69% correspond to the camera PAMPANEIRA 1 (i), 29.25% to PAMPANEIRA 2 (ii), 19.16% to BUBION (iii) and 25.9% to CAPILEIRA (iv) (see in Figure 4.2). We grouped the records based on the new vehicle identifier (num_plate_ID), taking into account the mobility behavior of each vehicle. For each vehicle, we built a record per each time the vehicle visits the area, containing the date of entry (entry_time_stamp) and exit (exit_time_stamp) to the area and a list of all the cameras (route) by which it has been registered during its stay, from which we can calculate the distance in kilometers traveled (total distance) in the area. From the above records, we can also calculate the duration of stay (avg_visit_POQ) expressed in days and the number of nights spent there. In case of missing data, i.e., we cannot calculate the time of entry or exit of a vehicle in the area, we remove the individual from the dataset.

After that, we performed a grouping at the license plate level so that each database row corresponds to a different individual. In this way, we fuse the information of all the vehicle visits in the area. Finally, we obtained a dataset with the total number of visits (total_entries), the average time (avg_visit_POQ) in days, the complete vehicle routing (route), the total accumulated distance traveled (total_distance), the standard deviation of the average time of each visit (std_visit_POQ) in days, the total time spent (total_time_POQ) in the area and the total number of nights spent there (nights). From the new record structure, we can calculate the visits of each vehicle in different weeks (visits_dif_weeks) and months (visits_dif_months) to study the fidelity of the individual in the area. Finally, we obtain a dataset with 50,901 vehicle records and ten attributes.

Phase	Tasks	Values
	Calendar Data	
Importing Data	Read the dataset with information on public holidays at national level in Spain	270 days, 3 attributes (date, type, holiday_period)
Set holiday periods	Establish the important holiday periods in Spain: Summer Holiday, Christmas and Holy Week	Summer Holiday (from 1 aug. to 31 aug.) Christmas (from 12 dec. to 6 jan.) Holy Week (from 10 apr. to 17 apr.)
Encode variables	Convert categorical holiday periods into binary variables	270 days, 5 attributes (date, type, Summer, Christmas, Holy_Week)
	License Plate Recognition Data	
Importing Data	Read the cleaned dataset produced from the detection of vehicle license plates	1,050,760 rows, 4 attributes (license_plate, time_stamp, type, camera_id)
Calculate associate variables	Calculate variables combining the 4 cameras	(license_plate, entry_time_stamp, exit_time_stamp, route, total_distance)
Group information	Group the information for each record by vehicle	50,901 rows, 10 attributes (license_plate, total_entries, avg_visit_POQ, std_visit_POQ, total_time_POQ, nights, route, total_distance, visits_dif_weeks, visits_dif_months)
	Vehicle information Data	
Importing Data	Reads the dataset with vehicle information and its origin	45,132 license plates, 4 attributes (license_plate, postcode, co2_emissions, num_seats)
	Demographic and Economic data	
Importing Data	Reads demographic information about the region of origin of the vehicle	11,752 regions, 4 attributes (postcode, population, gross_income, disposable_income)
Merging Data	Merge the two sources	INE
Validate Data	Validate information common to the two sources	INE
	Geographic data	
Importing Data	Reads information regarding the region of origin of the vehicle	11,752 regions, 8 attributes (postcode, autonomous_community, province, county, district, town, latitude, longitude)
Merging Data	Mix and validate information from the two sources used	geopy and pgeocode
Standardize values	Treatment of equivalences between names of regions in different co-official languages	Elimination of accents, spaces and translation to Spanish of all values related to region names
Validate Data	Validate postcodes and geolocation	geopy, pgeocode and INE
Generate new variables	Calculate the distance between the study area and the postcode origin from the coordinates	11,752 regions, 9 attributes (postcode, autonomous_community, province, county, district, town, latitude, longitude, km_to_POQ)
	Fusion Dataset	
Merging Data	Unification of header names and data formats, Mix postcode and license plate fields, Delete rows with some null fields	49,224 vehicles, 24 attributes (license_plate, total_entries, avg_visit_POQ, std_visit_POQ, total_time_POQ, nights, route, total_distance, visits_dif_weeks, visits_dif_months, co2_emissions, num_seats, postcode, autonomous_community, province, county, district, town, latitude, longitude, km_to_POQ, population, gross_income, disposable_income)
Generate new variables	Calculate variables related to the type of dates in the calendar during the period of stay of each vehicle	49,224 vehicles, 29 attributes (license_plate, total_entries, avg_visit_POQ, std_visit_POQ, total_time_POQ, nights, route, total_distance, visits_dif_weeks, visits_dif_months, co2_emissions, num_seats, postcode, autonomous_community, province, county, district, town, latitude, longitude, km_to_POQ, population, gross_income, disposable_income, total_holiday, total_workday, entry_in_holiday, total_high_season, total_low_season)
Exporting Data	Obtaining the resultant dataset	CLUSTERING_VEHICLES DB

TABLE 4.2: Detailed schematic of the data fusion stage in the pipeline.

Vehicle Information Data

The National Department of Traffic in Spain (DGT) has provided us with data relating to vehicle information[1] including details such as the vehicle's CO2 emissions (co2_emissions), the number of seats (num_seats) and the postcode of the vehicle's address (postcode). It is important to note that each vehicle is associated with a fiscal address used to pay road tax. This will generally match the place of origin of the vehicle driver, although as we have described in Section 3.6, this is not entirely true. This dataset will help us understand the distribution of vehicle types and ownership in the different regions. We have a dataset with 45,132 vehicles registered in Spain and four attributes. Unfortunately, we do not have this information for vehicles registered outside of Spain. The percentage of foreigners in the data sample is less than 9.5%. Therefore, we determined these individuals exclusively by their mobility behavior in the area. All information related to vehicle information, demographic, economic, and calendar holidays is restricted to Spanish-registered vehicles.

Demographic and Economic data

We have access to data regarding population size (population), average gross income (gross_income), and average disposable income (disposable_income) per person for each region linked to a postcode (postcode). This information comes from different sources such as the Spanish Tax Agency (AEAT)[2] or the National Statistics Institute (INE)[3]. The data are available for regions with more than 1000 inhabitants and are updated until 2020. When using the AEAT source, we noted that some regions were missing, so we relied solely on the INE as a source for the variables mentioned. The information collected in this database allows us to understand each region's economic and demographic characteristics, which can be valuable for analyzing patterns in the data related to the drivers' economic capacity and willingness to travel. We obtained a database with 11,752 postcode records from Spain and four attributes.

National calendar data

We obtain the holiday data using a holiday library, which also allows the creation of custom calendars for local holidays, long weekends, and bank holidays. The library is designed to quickly and efficiently generate holiday sets specific to each country and subdivision (such as state or province)[4]. It aims to determine whether a particular date is a public holiday and to set national and regional holidays for multiple countries. As we mentioned before, due to the small percentage of foreign individuals in the sample, and the complexity of dealing with a different set of holidays for different vehicles, we have restricted the analysis of the holidays to Spain. However, in the holidays, we include Saturdays and Sundays, so we also consider the idea of a weekly holiday for any origin. For each day, represented by a date (date), we have information (type) on whether it is a holiday or a working day in Spain. In addition, holiday periods have been defined to establish high and low tourist seasons based on the three most important national holidays in Spain: Summer, Christmas, and Holy Week[5], which represent a binary variable, indicating whether the date belongs

[1] https://sede.dgt.gob.es/es/vehiculos/informe-de-vehiculo/
[2] https://sede.agenciatributaria.gob.es/Sede/estadisticas.html
[3] https://www.ine.es/dynt3/inebase/es/index.htm?padre=7132&capsel=5693
[4] https://python-holidays.readthedocs.io/en/latest/
[5] https://es.statista.com/temas/3585/vacaciones-en-espana/#topicOverview

to that holiday period (Summer, Christmas, Holy Week). We obtain a database with 270 days and five attributes.

Geographic data

We get the geographic origin of the vehicles using the postcode and two libraries: pgeocode and geopy. pgeocode[6] allows fast and efficient queries of GPS coordinates, region name, and municipality name from postcodes. The library can also calculate the distances between postcodes. geopy[7] is a Python client that provides access to several popular geocoding web services. It allows developers to find the coordinates of addresses, cities, countries, and landmarks from third-party geocoders and other data sources. The library includes geocoding classes for numerous services, such as OpenStreetMap Nominatim and Google Geocoding API (V3), which can be found at geopy.geocoders. We use data from both sources to validate and complement each other's vehicle location information at different levels, such as municipality, county, or suburb. Furthermore, we also use data from the National Statistics Institute (INE)[8] to verify the province and autonomous community code of the vehicle, which is directly related to the postcode. Hence, we have created a database that contains, for each postcode, information about: (autonomous_community), (province), (county), (district), (town), (latitude), (longitude), and distance in kilometers between the origin of the vehicle and the study area (km_to_POQ). We obtain a database with 11,752 postal code records for Spain and nine attributes.

Merge of all the processed datasets

Finally, we fuse all constructed databases, crossing the information from the license plate and postcode variables. After merging the tables, we eliminated records with any of the aforementioned attributes null. The information from the national calendar allows us to add to the vehicle database information related to the stay and its total number of holidays (total_holiday), workdays (total_workday), high season (total_high_season), low season (total_high_season) and a binary variable indicating whether the vehicle enters the area on a holiday or a workday (entry_in_holiday). The resulting dataset contains information on the behavior in the area for 49,224 vehicles and 29 attributes.

4.4 Preprocessing

Our dataset contains about 29 attributes with different scales and units. Hence, some variables may be more influential than others in our analysis. To solve this problem, we will apply normalization to the data. Normalization must be applied to numerical data, so we must first convert the categorical variables to numerical values. Of the 29 variables, only the following are categorical: postcode, route, autonomous_community, province, county, district, town, license_plate. The variable license_plate is a unique identifier of each record, so we delete it. The region variable directly correlates with the km_to_POQ, so we can eliminate this feature without losing information. For the same reason, we eliminated the variables latitude and longitude. The numeric variable, total_distance, keeps the information of the kilometers traveled in the variable route. We will not use in this problem the variables

[6] https://pgeocode.readthedocs.io/en/latest/

[7] https://geopy.readthedocs.io/en/latest/

[8] https://www.ine.es/daco/daco42/codmun/cod_ccaa_provincia.htm

co2_emissions and num_seats, since they contain empty values for a high percentage of individuals (approx. 25%). Also, we believe that for this problem, they do not provide information of relevant interest. We will finally obtain a dataset with 49,224 vehicles and 17 numerical attributes: total_entries, avg_visit_POQ, std_visit_POQ, total_time_POQ, nights, total_distance, visits_dif_weeks, visits_dif_months, km_to_POQ, population, gross_income, disposable_income, total_holiday, total_workday, entry_in_holiday, total_high_season, total_low_season.

4.5 Dimensionality reduction

It is advisable to reduce the search space of the feature matrix before the clustering for efficiency reasons. This process aims to create a simplified set of dimensions containing much of the variance of the original dataset, as there could be features with very low variance, which would make them of little use for our goal of clustering into different vehicle behaviors. Specifically, we want to minimize the number of components as a function of the variance explained. Hence, after the normalization in Section 4.4, we apply Principal Component Analysis (PCA), as a dimensionality reduction algorithm. The result of this stage is a projection of the original dimensions onto a smaller number of new dimensions, also called PCA components. Then, we will evaluate which normalization offers a higher variability value on a fixed value of PCA components. Visualizing the variance explained by each component can help choose the best normalization. This analysis will be discussed in Chapter 5.

Removing highly correlated features can cause a loss of information for data with high dimensionality, and the PCA technique reduces the dimension of the dataset [23]. However, we have found that removing variables with very high correlation substantially improves the results and the performance of the clustering models for our data. Furthermore, correlated variables increase the data's variance, making the visual interpretation of the PCA results difficult, as the firsts principal components may not accurately reflect the underlying structure of the data.

4.6 Clustering and evaluation

Our choice of unsupervised machine learning is motivated by the need to categorize unlabeled individuals into distinct groups. Our study explores all the algorithms mentioned in Section 3.1 to determine the optimal approach for pattern recognition and evaluate whether they can find a realistic solution.

4.7 Visualization

Data visualization is essential in our work, as it helps to determine and make decisions about parameter settings, algorithms, and normalization used in our analysis. Likewise, it helps in the explainability of the machine learning results. Regardless of the metrics used, performing a visual evaluation of the results is recommended. Combining the evaluation metrics and visualization charts, we can reach conclusions that facilitate making informed decisions on configuring our model. For example, we use the elbow method to determine the optimal number of clusters for different algorithms. This method consists of plotting a graph based on the number of clusters or components and a given evaluation metric, and visually looking for

the inflection point on the defined curve (“elbow"). By selecting the number of components at the bend, we strike a balance between model complexity and accuracy.

We have visualized the first two PCA components for each normalization with scatter plots to study the data’s geometry and the clusters’ distribution. These visualizations help to understand the underlying structure of the data. However, there is a limitation in the number of coordinates supported for the scatter plot (maximum 3). Some graphs, such as the Andrews curves, exceed the theoretical limit of the number of components displayed, as they are based on the Fourier series. The problem arises because too many individuals in the dataset overlap, which causes the cluster curves to overlap as well and, thus, provide limited visual information. Therefore, this graph is not helpful for our data. We also use box plots to visualize the distribution of each feature among the individuals in each generated cluster. By comparing these distributions between different clusters, we can identify common patterns that help us understand the individuals that compose each cluster.

Chapter 5

Results

The first intuitive analysis is a correlation to identifying possible predictors of residency. As we want to model the behavior of the traffic, and we assume that the behavior of the residents should be different from the tourists' behavior, we correlate the residence label against all the other variables. We assign a label of 1 if the vehicle is registered in Pampaneira, Capileira or Bubión (registered resident) and 0 otherwise. Our results reveal several variables that showed non-significant correlations (correlation less than 0.2): avg_visit_POQ, std_visit_POQ, and population, which have been removed (see in Figure 5.1). Table 5.1 shows the mean and standard deviation for a selection of uncorrelated variables between two groups: registered residents and non-registered individuals. The variable visit_POQ, which relates to the time of each visit to the area, has a high standard deviation (std_visit_POQ) for registered residents, which could indicate the existence of residents who go out more often and others less often. Furthermore, the mean values (avg_visit_POQ) of this variable (visit_POQ) do not differ much between the two groups. As for the population size variable, there are other villages with a similar size to that of the three villages under study. Hence, the three removed variables are less relevant than the rest in predicting whether a vehicle is a resident of the study area or not.

Preprocessing and Dimension reduction results: Normalization selection

Although the preprocessing and dimension reduction stages are performed sequentially, they are interdependent, so we will describe them together.

	nights		total_distance		total_entries		entry_in_holiday	
	Residents	Others	Residents	Others	Residents	Others	Residents	Others
mean	158.47	19.99	205.82	13.60	19.46	2.35	4.26	0.72
std	72.37	48.07	238.52	47.78	23.57	6.58	5.49	1.70
	gross_income		km_to_POQ		visits_dif_weeks		total_high_season	
	Residents	Others	Residents	Others	Residents	Others	Residents	Others
mean	16,084	25,007.07	1.02	374.73	4.57	1.48	27.53	3.84
std	0.00	7671.19	0.59	486.97	4.03	1.97	14.75	9.00
	total_holiday		avg_visit_POQ		std_visit_POQ		population	
	Residents	Others	Residents	Others	Residents	Others	Residents	Others
mean	52.54	6.83	23.60	10.54	20.26	4.15	406.66	19,8175.90
std	23.71	15.06	34.85	31.87	23.35	16.05	121.16	56,7183.30

TABLE 5.1: Mean and std. deviation for registered residents and rest of individuals in dataset.

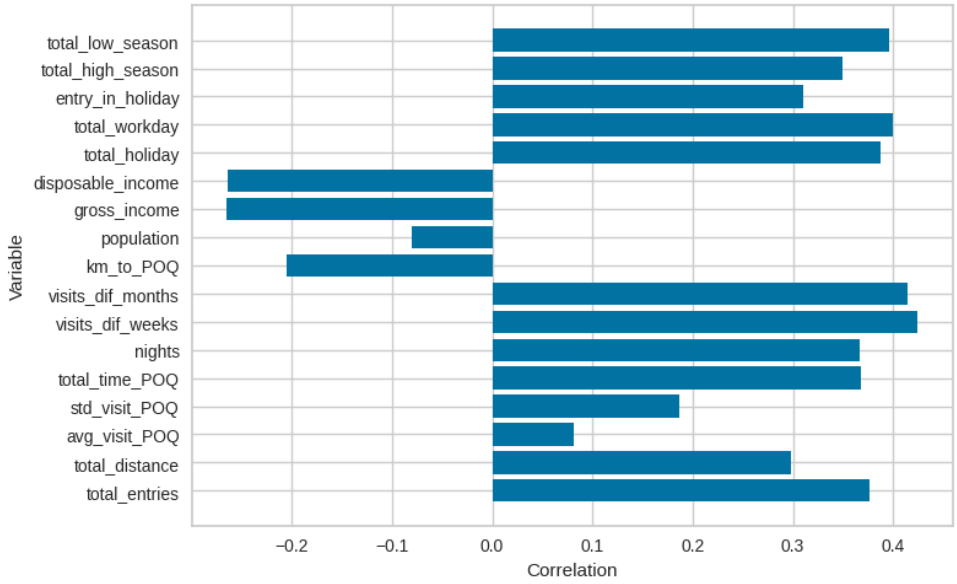


FIGURE 5.1: Correlation between the registered resident label and the rest of the variables.

PCA technique handles the linear correlations of the original variables to create the principal components. However, we observed that the prior removal of highly correlated variables improves the variance explained by PCA and the scatter plots displaying the PCA component. Hence, we plot in Figure 5.2 the correlation matrix, in which each entry represents the correlation coefficient between a pair of attributes in the dataset. We identify five pairs of variables that are highly correlated, which may cause multicollinearity problems in the study. Specifically, we remove variables that have a correlation coefficient higher than 0.90 with other variables, reducing the attributes in our dataset. We therefore use the following variables: total_entries, nights, visit_dif_weeks, visit_dif_months, km_to_POQ, gross_income, entry_in_holiday, total_distance and total_high_season.

After applying the four most common normalization to the data (see in Chapter 3), we apply PCA analysis. Figure 5.3 shows the variance carried by each PCA component for each normalization. We can appreciate that two components explain most of the variance in all normalization. Hence, we perform an exploratory visual analysis plotting the first two principal components to study their underlying geometry. We have overlapped on the plots, in red, the points representing the vehicles of the registered residents. These visualizations provide information about the data structures and the performance of each normalization method. Based on the principal component analysis and the visualization of the data geometry, we will choose which type of algorithm and normalization is the most suitable for our problem (see in Figure 5.4).

The normalization method that obtained the highest cumulative variance is ℓ^2, indicating that it retains the most information in only two components (see in Figure 5.3 (d)). In addition, the variance of each dimension is high compared to the other techniques analyzed, suggesting that the data are well distributed in both dimensions. The graph in Figure 5.4 (d) shows a clear separation between the two groups, and the registered residents (in red) are well confined. The min-max normalization method obtained the second-best cumulative variance and the highest

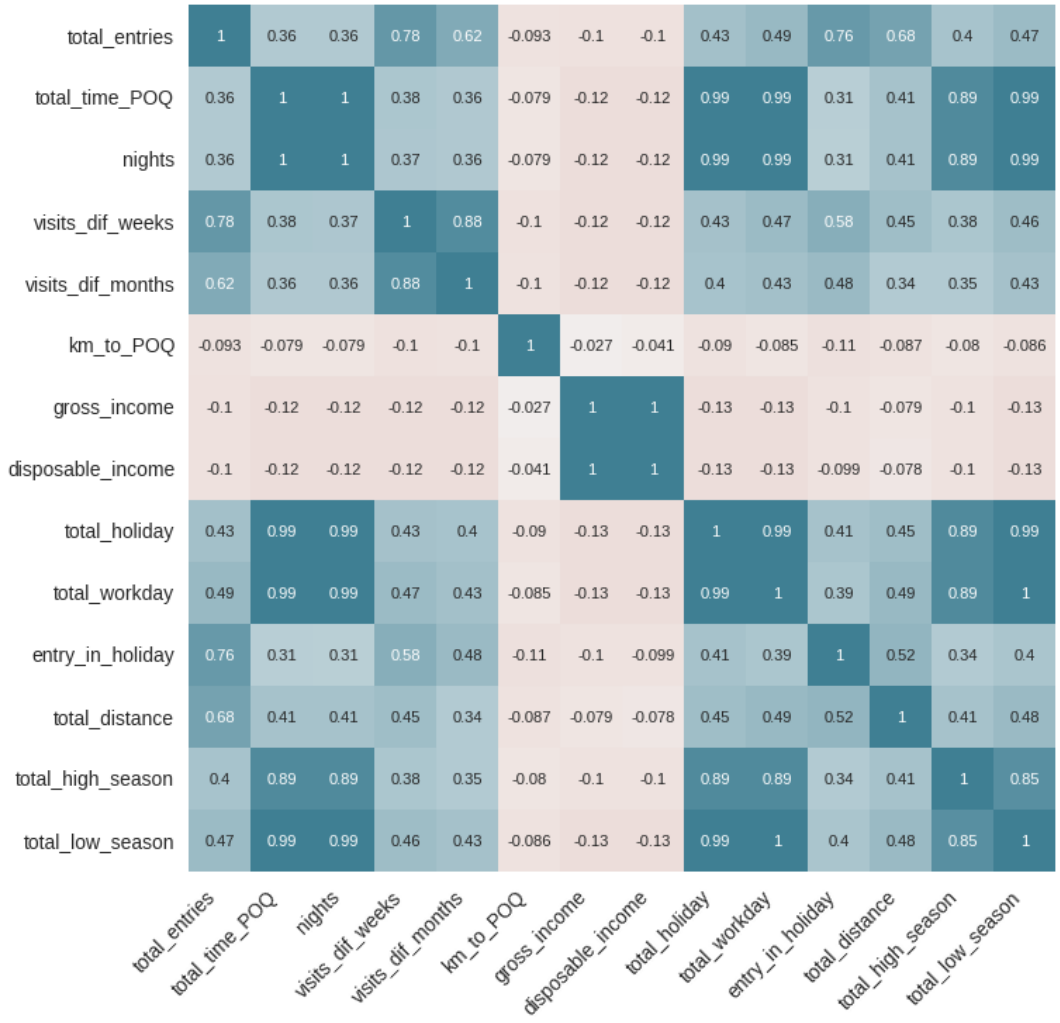
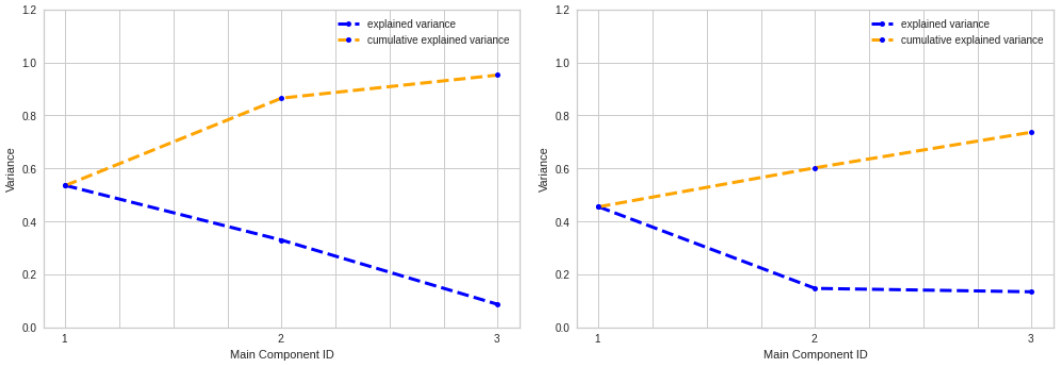
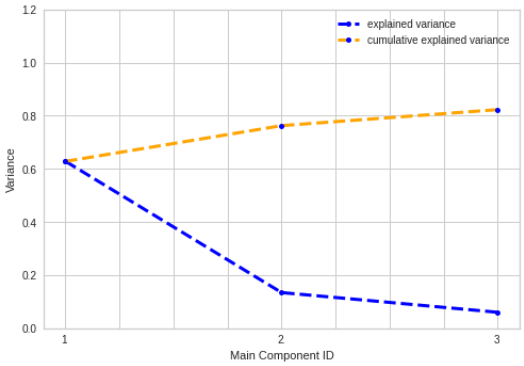


FIGURE 5.2: Correlation matrix for all variables in the proposed dataset.

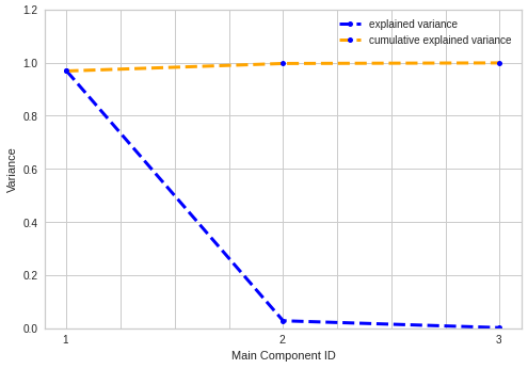


(a) Min-max normalization.

(b) Z-score standarization.

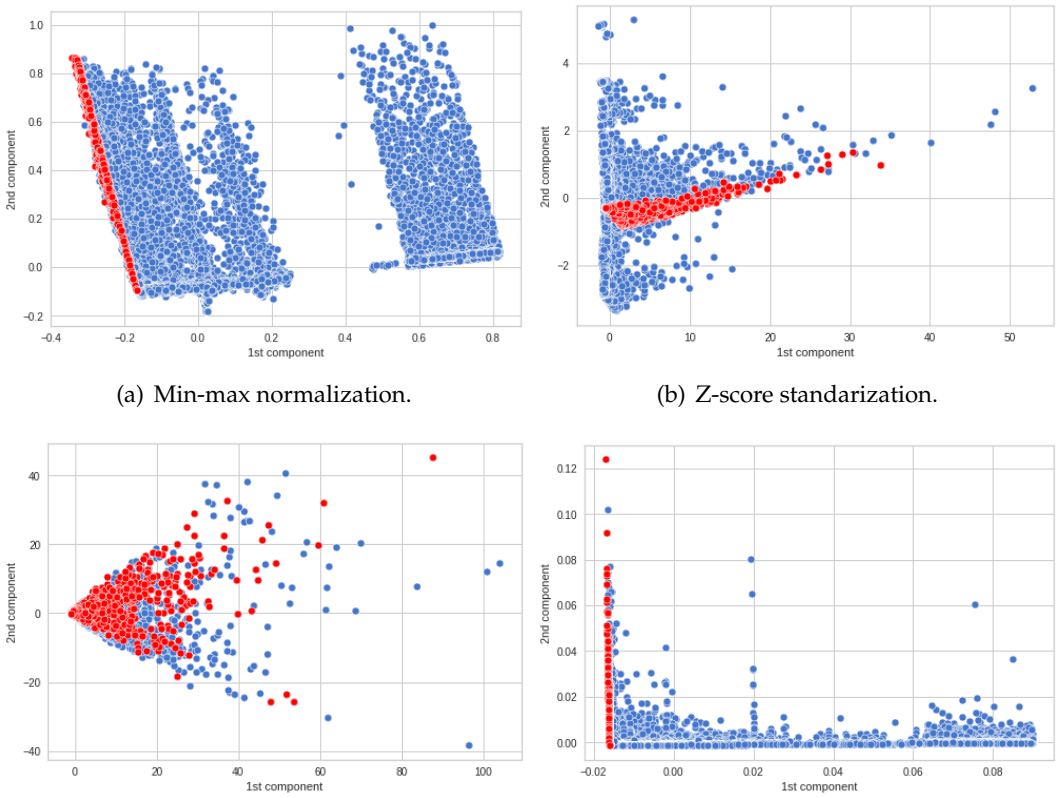


(c) MAD normalization.



(d) ℓ^2 normalization.

FIGURE 5.3: Variance with 3 principal components.



(a) Min-max normalization.

(b) Z-score standarization.

(c) MAD normalization.

(d) ℓ^2 normalization.

FIGURE 5.4: Scatter-plot of the first two principal components for the different normalizations.

variance for each dimension, preserving a reasonable amount of information in only two components (see in Figure 5.3 (a)). The graph also shows a clear separation between the two groups, and the actual residents are defined along a vertical line on the left cluster in Figure 5.4 (a). In contrast, the mad normalization method has a lower cumulative variance and variance for each dimension (see in Figure 5.3 (c)) than the ℓ^2 and min-max normalization methods. The 2-dimensional scatter plot shows no apparent clusters (see in Figure 5.4 (c)), and the actual residents are highly dispersed, which makes it unusable for our analysis. We had similar results in a scatter plot of three principal components. Finally, the mean normalization, z-score, method presented the lowest cumulative variance, indicating that it loses more information during dimensionality reduction than other techniques (see in Figure 5.3 (b)). The graph shows that the actual residents are grouped together, but for the 2-components, there are no apparent significant clusters (see in Figure 5.4 (c)). The trend of the cumulative variance explained is rising, suggesting that the current normalization method could be enhanced by including more components. By adding more dimensions, it may be possible to identify a dimension where the group of registered residents conforms to a clearer distribution. Principal Component Analysis (PCA) typically works better with z-score standardization than with min-max normalization. However, normalization techniques that better handle outliers (such as z-score) may not always be effective for all datasets because it tries to distribute the individuals uniformly, softening the outliers. For example, we observed that the min-max normalization method performed better than the z-score standardization, possibly due to the presence of small clusters that z-score detects as outliers. In particular, the dataset has a low proportion of registered residents (less than 2% of the total sample), which could be considered outliers (see in Table 5.1). In these cases, the min-max normalization method, which is more sensitive to small clusters, may give better results. With all this information, we decided to apply the two best normalizations for our data (ℓ^2 and min-max) and compare the results obtained in the clustering.

From the scatter plots in Figure 5.4, we observe that the data points are spread relatively flat. This suggests that the data points are concentrated in a lower dimensional space within the original feature space. In other words, the data appears to exist in a more compressed space rather than being spread out across multiple dimensions. Hence, partition and distribution-based clustering models are the most suitable for this geometry (see in Section 3.5). To verify this, we have also tested other algorithms with poor results. For example, density and spectral-based algorithms performed poorly, probably because of the non-flat geometry, but also because they work best for detecting outliers. Hierarchical algorithms performed poorly, probably because of the non-flat geometry, but also they have difficulties with highly concentrated datasets, creating distinct groups only when the separation is very obvious. Consequently, we decided to focus on the partitional and distribution-based algorithms, which work well with flat geometry data. In particular, we try Gaussian mixtures, K-Means, and MiniBatchKMeans.

Gaussian Mixture models are more flexible and can handle different cluster shapes and sizes, while K-Means assumes a spherical shape of the clusters and a uniform size. In addition, Gaussian Mixture models can estimate the probability that a data point belongs to a cluster, which can be useful in specific applications where we need to make decisions based on uncertain data or when we want to assign a data point to multiple clusters with different probabilities. On the tests carried out, we discovered

that K-Means and MiniBatchKMeans are not able to find any cluster that contains the majority of individuals of registered residents (see in Figure 5.4 (a) and (d)). This is because the distribution of these individuals follows an elliptical geometry, which is not amenable to partition-based algorithms directly. Based on these results, we used the Gaussian Mixture clustering algorithm given the geometry of our data and the distribution followed by registered residents.

Evaluation results

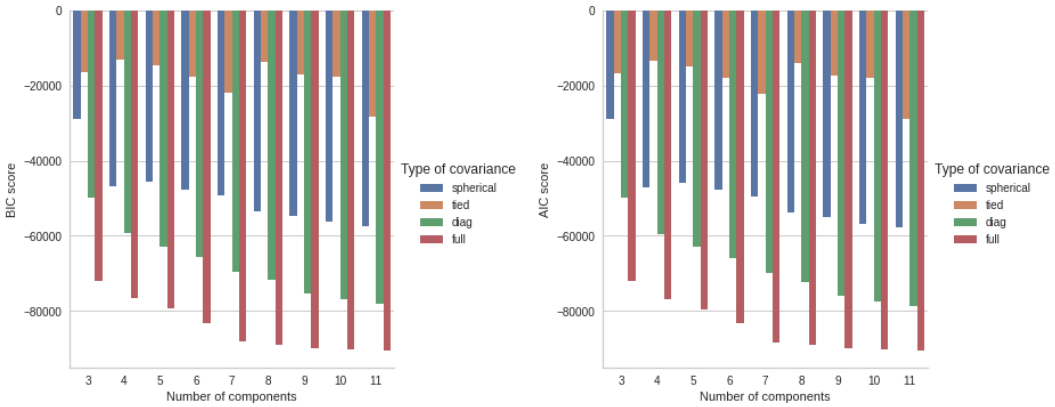
Once selected the algorithm, we need to select the configuration and hyperparameters of the algorithm. In GaussianMixture algorithm, a "mixture" refers to a linear combination of several Gaussian distributions, where each component represents a Gaussian distribution in the mixture [56]. Each mixture component is defined by a set of parameters, including its own mean and covariance matrix, which describes how the data are distributed in that space. In practice, the number of mixture components is an adjustable parameter of the algorithm, meaning that we can specify how many Gaussian distributions to combine to model the data. Another configurable parameter of the GaussianMixture algorithm covariance type used to construct the covariance matrix associated with each mixture component. These covariance types specify how the different variables in the data are correlated and can significantly affect the accuracy and efficiency of the model. The common types of covariance are:

- Full: all components have their own covariance matrix. This means that each component can have a complex correlation structure between the different variables.
- Tied: all components share the same overall covariance matrix. This can be useful if different variables are highly correlated.
- Diagonal: each component has its own diagonal in the covariance matrix. This means that the correlation structure between the different variables is limited to correlations between pairs of variables.
- Spherical: each mixture component has its own unique variance. This means that the correlation structure between the different variables is limited to the variance of each variable individually.

To select the best hyperparameters, we calculate the performance of the resulting model with the metrics presented in Section 3.2, which are appropriated for clustering algorithms based on density (BIC and AIC). In the next subsections, we perform the evaluation for the different types of covariance of the GaussianMixture algorithm on the two normalizations chosen in the previous subsection: min-max and ℓ^2 normalization.

Evaluation results: Min-max normalization

We begin first by presenting the results obtained with the min-max normalization. Figure 5.5 represents the values of the BIC and AIC metrics with respect to the number of components and type of covariance used as parameters of the GaussianMixture algorithm. We note that the "full" covariance type is the one that minimizes both metrics in all cases, so it will be the one chosen for the subsequent analysis. This value means that each component has its own overall covariance matrix, which



(a) BIC score. (b) AIC score.

FIGURE 5.5: Information criteria for the GaussianMixture on min-max normalization.

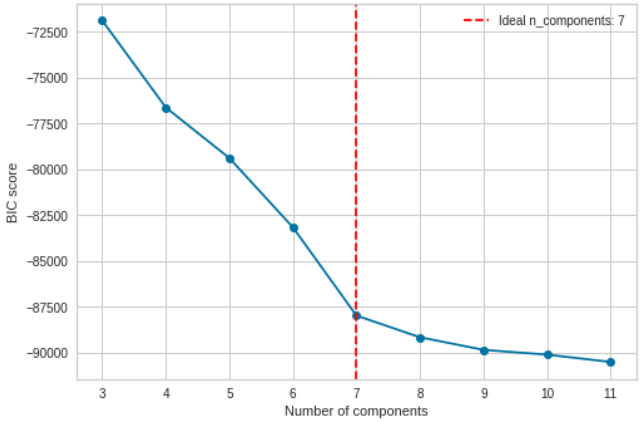
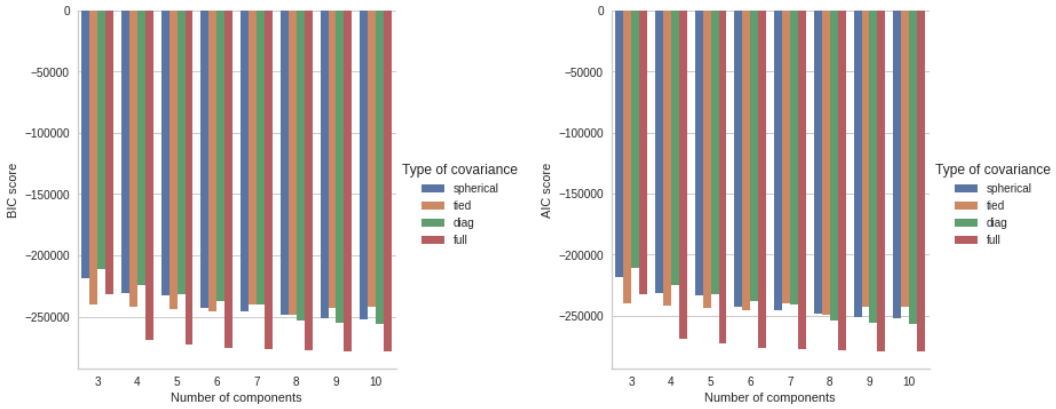


FIGURE 5.6: Elbow method for BIC using min-max normalization.

means it can capture any correlation between variables. We note no significant differences between the values obtained for AIC and BIC scores. Therefore, we calculate the elbow method on the BIC score to select the optimal number of mixture components, which from Figure 5.6 is 7, producing an abrupt change in the slope of the curve.

Evaluation results: ℓ^2 normalization

Figure 5.7 represents the values of the BIC and AIC metrics with respect to the number of components and type of covariance, used as parameters of the GaussianMixture algorithm for ℓ^2 normalization. We observe that the "tied" covariance type is slightly superior for 3 components, although for more than 3 components the "full" covariance type is again the best. Similarly to the min-max normalization, there is no significant difference between the values obtained for AIC and BIC scores. Therefore, we will calculate the elbow method on the BIC score (see in Figure 5.7) and "full" covariance type. The elbow method indicates that there is a sharp change in the slope at 4 components.



(a) BIC score. (b) AIC score.

FIGURE 5.7: Information criteria for the GaussianMixture on ℓ^2 normalization.

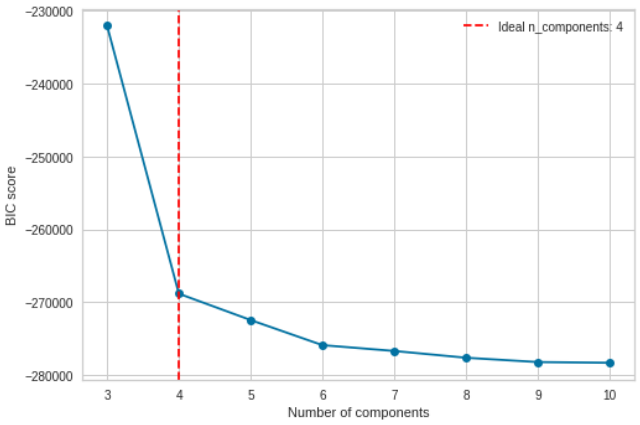
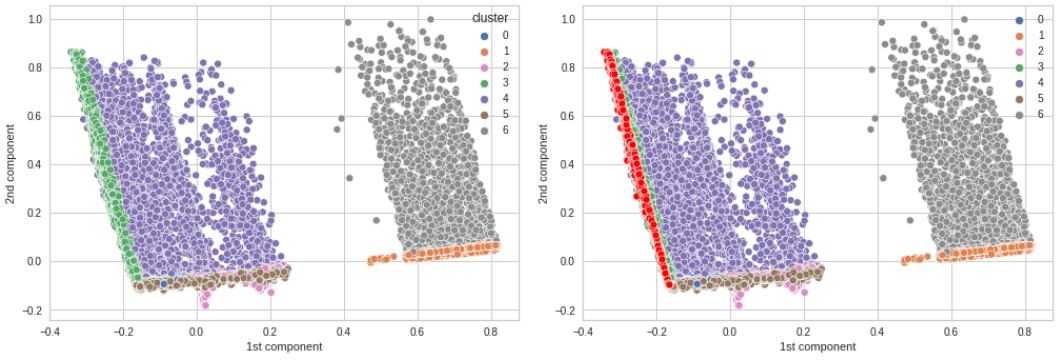
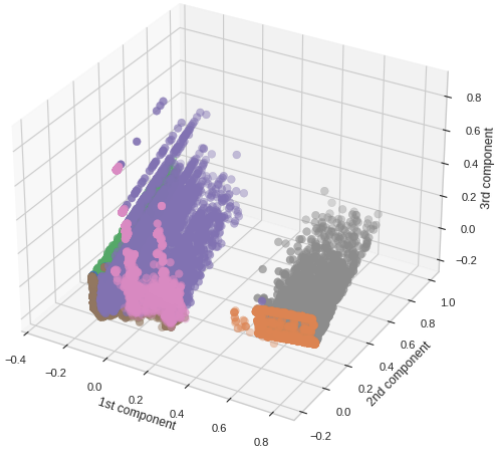


FIGURE 5.8: Elbow method for BIC using ℓ^2 normalization.



(a) Segmentation for 7 mixture components.

(b) Highlighted registered residents.



(c) 3D plot with 3rd component.

FIGURE 5.9: Scatter-plot of the first three components (PCA) using min-max normalization.

Visualization results

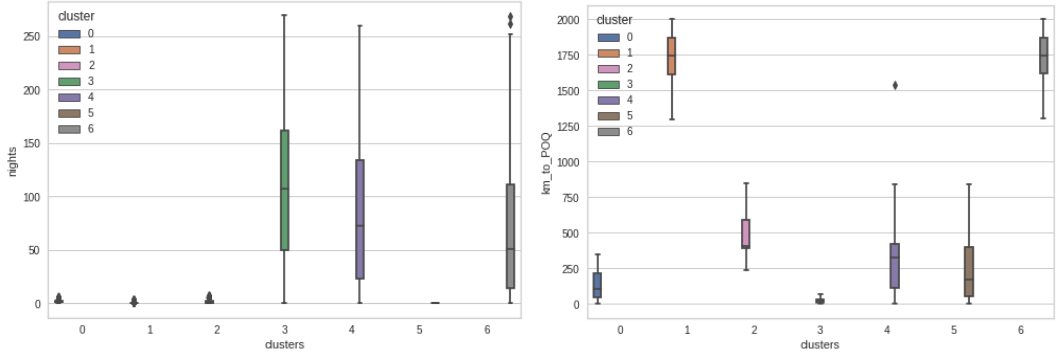
Once we have selected the clustering algorithm and the hyperparameters, we will discuss the visualization of the generated clusters over the two chosen normalizations: min-max normalization and ℓ^2.

Visualization: Min-max normalization

Figure 5.9 (a) shows a 2D scatter plot, where each axis represents one of the principal components (1st and 2nd) of the distribution. In Figure 5.9 (b), we highlight in red the registered residents labeled in our database. Finally, in Figure 5.9 (c), we can visualize the 3D scatter plot, in which each axis represents one of the 3 principal components. Table 5.2, shows the percentage of vehicles and the number of registered residents in each of the 7 clusters. We can see that cluster 3 correctly groups more than 96% of this group of individuals. Approximately 48% of the total sample is grouped in only one cluster (cluster 5), so almost half of the vehicles follow the same behavior. In addition, the cluster containing most of the registered residents (cluster 3) represents approximately 11% of the total population.

Data points	N° cluster						
	0	1	2	3	4	5	6
Percentage of sample	15.04%	6.11%	8.58%	11.17%	8.55%	47.50%	3.05%
Real Residents	10	0	0	641	3	12	0
Rest of individuals	7391	3009	4221	4862	4205	23,370	1500

TABLE 5.2: Clusters based on registered resident labels using min-max normalization.



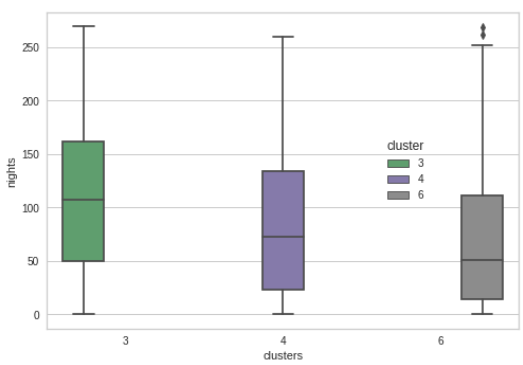
(a) Total nights of stay by cluster.

(b) Distance in Km. to the area by cluster.

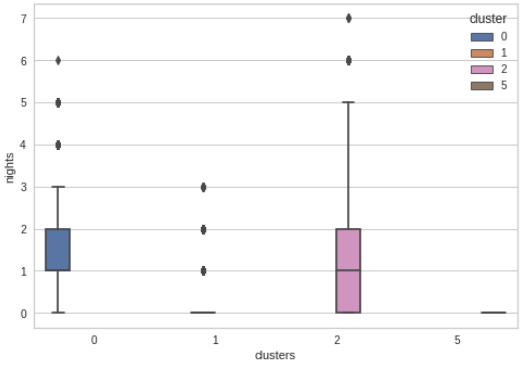
FIGURE 5.10: box plots for min-max normalization (I).

Figure 5.10 presents the box plots for the 7 clusters for the nights (Figure 5.10 (a)) and km_to_POQ (Figure 5.10 (b)) variables, which show significant differences in explaining the groups. Figures 5.11 and 5.12 present the box plots of the most relevant variables for the 7 clusters obtained, making a previous division into 2 groups, which we explain below. Table 5.3 complements Figures 5.11 and 5.12, indicating the exact number of the mean of each variable in each cluster. To facilitate visualization, we have separated some of the box plots according to the value of the variable nights, which seems to discriminate well between 2 groups of clusters: (0, 1, 2, 5) with lower values and (3, 4, 6) with higher values (see in Figure 5.10 (a)). Clusters 3,4,6 have a number of nights close to the behavior of a resident in the area. These represent 22.77% of the total sample (see in Table 5.2). Clusters 0,1,2,5 have visitor behavior because they spent fewer nights in the area and represent 77.23% of the total dataset. Thus, we can define a distinction between the 7 clusters based on length of stay, which is directly correlated with overnights.

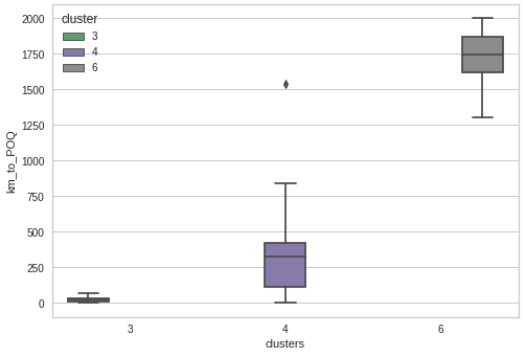
For clusters 3,4,6 (residents' behavior), another relevant variable is the distance in kilometers from the registered address of the vehicle to the area (see in Figure 5.11 (c)). The three clusters, despite having notable differences for the variable of origin, have similar behaviors for the variable of nights. Hence, the individuals in these 3 clusters have residence or accommodation in the area. According to their distance to the area, cluster 3, with a mean value of 19.39 km (see in Table 5.3), corresponds mostly to vehicles registered in the area under study (registered residents) and nearby villages in the Alpujarra. Cluster 6, with a mean value of 1747.30 km for the variable km_to_POQ, corresponds to non-registered residents from abroad that we defined in Section 3.6. Cluster 4, with a mean value of 318.36 km, corresponds to individuals from other regions of Spain who are non-registered residents, which we also discussed in that Section. Moreover, the gross income variable is much higher



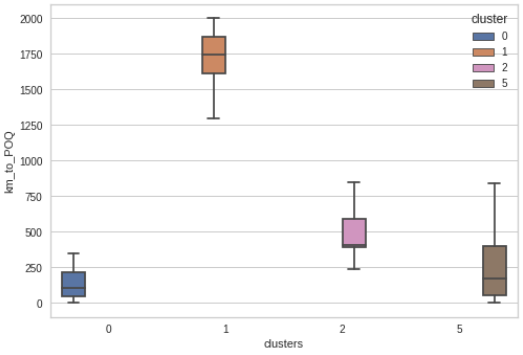
(a) Nights of stay for clusters **3,4,6**.



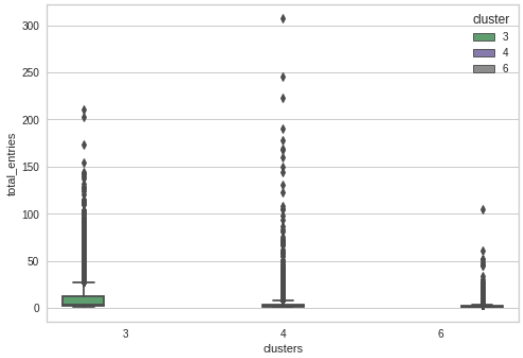
(b) Nights of stay for clusters **0,1,2,5**.



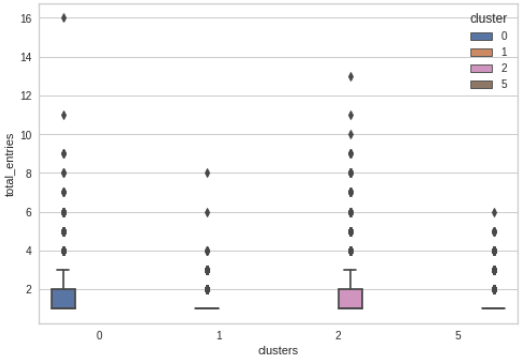
(c) Distance to the area for clusters **3,4,6**.



(d) Distance to the area for clusters **0,1,2,5**.

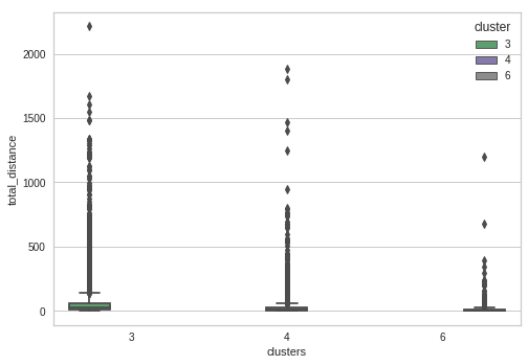


(e) Total entries for clusters **3,4,6**.

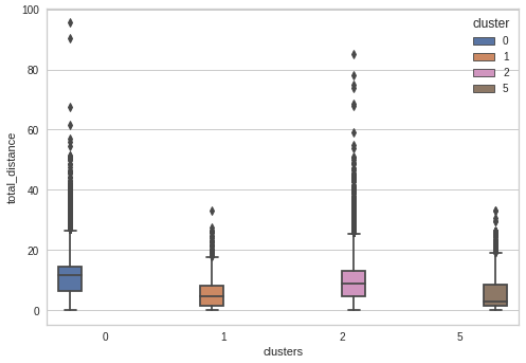


(f) Total entries for clusters **0,1,2,5**.

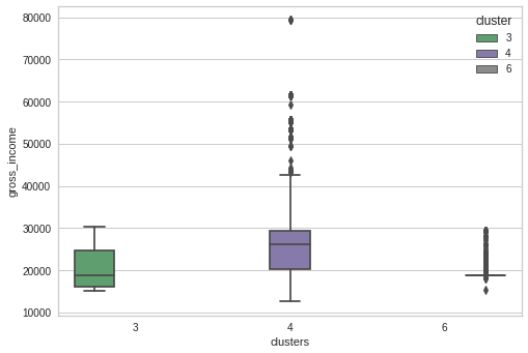
FIGURE 5.11: Box plots for min-max normalization (II).



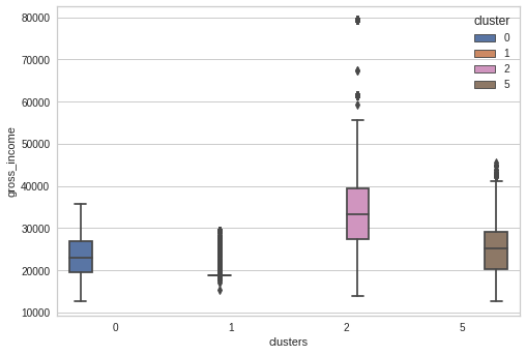
(a) Distance run in area for clusters **3,4,6**.



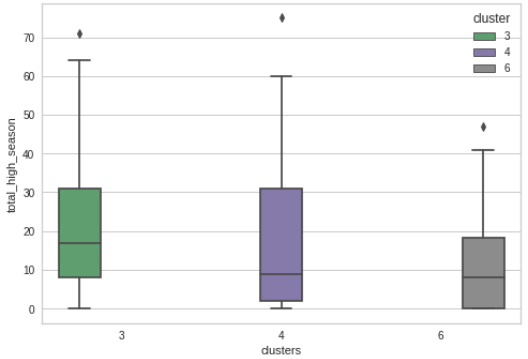
(b) Distance run in area for clusters **0,1,2,5**.



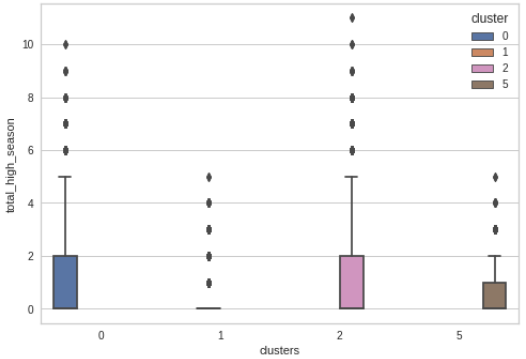
(c) Avg. gross income for clusters **3,4,6**.



(d) Avg. gross income for clusters **0,1,2,5**.



(e) Total high season for clusters **3,4,6**.



(f) Total high season for clusters **0,1,2,5**.

FIGURE 5.12: Box plots for min-max normalization (III).

in group 4 than in groups 3 and 6 (almost 34% higher) (see in Figure 5.12 (c)). Therefore, we can say that the majority of individuals in this group (non-registered residents from other Spanish regions) come from regions with higher incomes than the average of the rest of the residents. We can observe that in general for the three groups of residents, the mean of the variables total_distance, total_high_season, and total_entries (see in Table 5.3), are inversely proportional to the mean of the clusters for the variable km_to_POQ. Therefore, residents coming from farther away (clusters 4 and 6) will have a lower mean than the clusters coming from closer (cluster 3) for those variables, this is because coming from farther away the frequency of visitation, kilometers traveled and visits in high season (see in Figure 5.11 (e) and Figure 5.12 (a, e)) will also be lower.

For clusters 0,1,2,5 (visitor behavior), we also describe the average behavior of each cluster (see in Table 5.3). Cluster 0 has an average distance of 128.55 km to the area, so it corresponds to visitors from the province of Granada (region of the villages). This cluster comprises individuals who spend an average of 1.57 nights in the area. The variable total_entries tells us that individuals have made an average of 1.54 visits during the period collected in the sample. Due to its variables total_high_season (see in Figure 5.12 (b, f)), we observe that more than 65% of the visits are made in high season. Therefore, this cluster corresponds to individuals from the province of Granada who visit the area on weekends and holidays and stay between 1–2 nights. Cluster 1 has an average of 1742.97 km, which indicates foreign visitors. This cluster has an average of 0.26 nights, so they are individuals who usually visit the area during the day. Observing the variable total_high_season, we can see that the behavior of these individuals who come from abroad is to visit in low seasons. In addition, the relatively low value of the total_distance variable (4.90 km) suggests that it is likely that these tourists visit one of the three villages, specifically Pampaneira, using the main road instead of deviating from other routes, probably the behavior of these individuals is to visit the towns of the Alpujarra that coincide with the route of the main road. Cluster 2 has an average of 474.21 km, so visitors come from outside the province of Granada. This cluster comprises individuals who spend an average of 1.55 nights in the area, similar to cluster 0 (visitors from the province of Granada). This cluster has the highest average for the gross_income variable of all the clusters (see in Figure 5.12 (d)). This leads us to think that it contains vehicles from the north of Spain, where the average income is higher than in the south of Spain. In addition, the variable total_high_season, shows that approximately 74% of the stays are in high season, so these individuals correspond to tourists from the northern regions of Spain who decide to spend 1–2 days in the Alpujarra during their holidays. Finally, cluster 5 has an average of 253.70 km, indicating visitors from other Andalusia provinces. This cluster has an average of 0 nights, so the visits usually occur during the day, with no associated overnight stays. It is important to highlight that cluster 5 represents 47.50% of the individuals in the sample (see in Table 5.2), so we can say that the majority behavior among individuals is not to spend the night in the area. Furthermore, they rarely come in high season (27% of the total entries) (see in Figure 5.12 (f)), so the majority will correspond to tourists from other provinces of Andalusia who visit the area and return home at the end of the day.

Variables	Nº cluster						
	0	1	2	3	4	5	6
nights	1.57	0.26	1.55	108.62	84.66	0.00	68.73
km_to_POQ	128.55	1742.97	474.21	19.39	318.36	253.70	1747.30
total_entries	1.54	1.12	1.58	10.34	4.36	1.12	2.71
total_distance	11.64	4.90	10.67	70.24	30.77	4.86	14.42
gross_income	23,085.36	19,482.10	35,547.66	20,972.17	26,902.26	25,151.75	19,179.54
total_high_season	1.01	0.31	1.14	18.85	15.10	0.31	11.24

TABLE 5.3: Mean of variables for each cluster performed using min-max normalization.

Data points	Nº cluster			
	0	1	2	3
Percentage of sample	8.55%	8.76%	4.36%	78.33%
Real Residents	589	0	0	77
Rest of individuals	3620	4314	2146	38,478

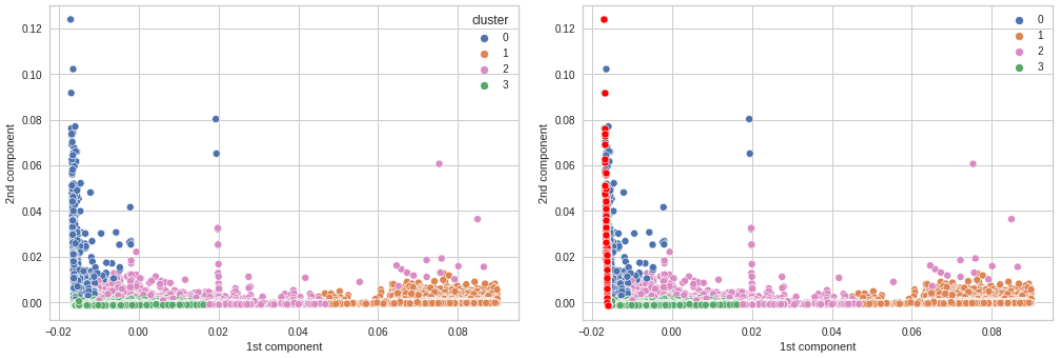
TABLE 5.4: Clusters based on actual resident labels using ℓ^2 normalization.

Visualization: ℓ^2 normalization

Figure 5.13 shows the distribution of the data using the ℓ^2 normalization. Figure 5.13 (a) shows a 2D scatter plot, where each axis represents one of the principal components (1st and 2nd) of the distribution of the groups. In Figure 5.13 (b), we highlight in red the registered residents. Finally, in Figure 5.13 (c), we can visualize the 3D scatter plot. Table 5.4 shows the percentage of the total sample and the number of registered residents in each cluster. We can see cluster 0 correctly groups more than 88% of the registered residents. Approximately 80% of the total sample is grouped in a single cluster (cluster 3). In addition, the cluster containing registered residents (cluster 0) represents approximately 9% of the total population.

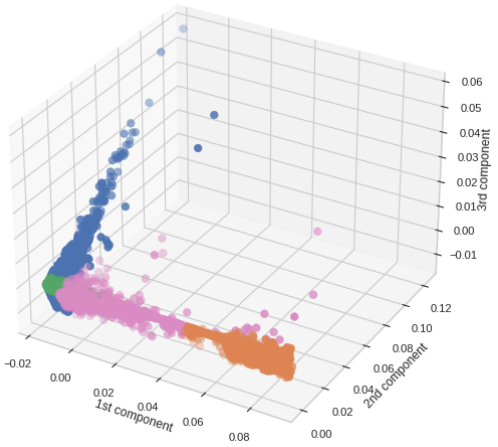
Figure 5.14 shows the box plots of the relevant variables for the 4 clusters, and Table 5.5 shows the mean of each of these variables in each cluster. We distinguish two clusters that contain a high value of the variable nights (cluster 0 and 2), while the rest of the clusters (clusters 1 and 3) have a low value. Although there are outliers (see in Figure 5.14 (a)) that increase the mean number of nights for these clusters (clusters 1 and 3), 50% of the individuals have a number of nights lower than 2 for cluster 3 and lower than 15 nights for cluster 1.

Cluster 0 has an average of 144.93 nights and an average distance to the area of 25.54 km. In addition, it contains more than 88% of the real residents. hence, we can consider this cluster as that of the area's residents (registered and unregistered). Most of the unregistered residents in this group, as shown in Figure 5.14 (b), come from the province of Granada. Cluster 2 is small (only 4.36% of the total sample) of non-registered residents who spend an average of 84.62 nights and come from an average origin of 598.01 km. Therefore, it corresponds to residents from outside the province of Granada. We can observe that in general, for the two groups, the means of the variables total_distance, total_high_season and total_entries (see in Table 5.5), are inversely proportional to the means of the variable km_to_POQ. This means that visitors that come from further away tend to: visit in the low season; move less inside the area; and come fewer times in the year than other visitors (see in Figure 5.14 (c, d, f)). Cluster 1 comprises individuals from afar to the 1750.68 km



(a) Segmentation for 4 mixture components.

(b) Highlighted registered residents.



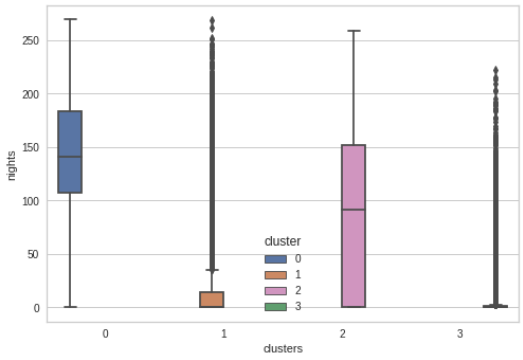
(c) 3D plot with 3rd component.

FIGURE 5.13: Scatter-plot of the first three components (PCA) using ℓ^2 normalization.

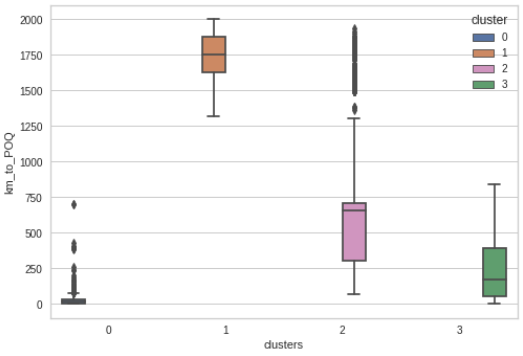
Variables	Nº cluster			
	0	1	2	3
nights	144.93	22.81	84.62	4.82
km_to_POQ	25.54	1750.68	598.01	240.01
total_entries	13.18	1.52	3.53	1.49
total_distance	95.43	6.89	26.43	8.01
gross_income	20,268.41	19,018.55	22,886.88	26,158.32
total_high_season	24.83	3.87	14.47	1.36

TABLE 5.5: Mean of variables for each cluster performed using ℓ^2 normalization.

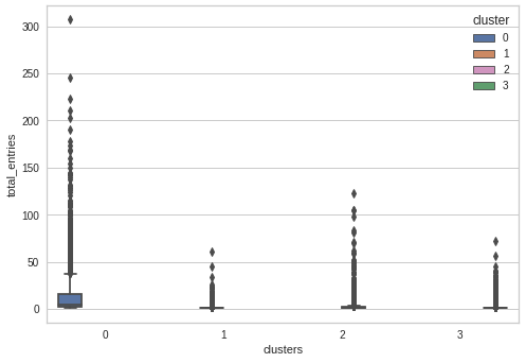
area and has an average of 22.81 nights (although most of them spend less than two nights). This group contains foreign visitors and some unregistered foreign residents (less than half of the group). Only 17% of the stays are in high season. This is because the cluster comprises foreigners, so they do not depend on the Spanish national calendar. Furthermore, despite having an average of 22.81 nights, the visits to the area are only 1.52, which means that in all those days they do not travel to nearby areas (see in Table 5.5). Finally, group 3 contains the majority of individuals in the data set and models individuals with a mean behavior of 4.82 nights (although most do not stay overnight) and 240.01 km of distance. This cluster represents 78.33% of the individuals in the sample (see in Table 5.4). Hence, we can say that the majority behavior of the individuals is not to spend the night in the area. In addition, they rarely come in high season (28% of the total stay) (see in Figure 5.14 (f)). Thus, the majority will correspond to tourists from Granada and other nearby Andalusian provinces visiting the area and returning home at the end of the day. Cluster 3 has the highest income, with a mean of 26,158.32; however, since it contains almost 80% of the sample, it does not provide much information.



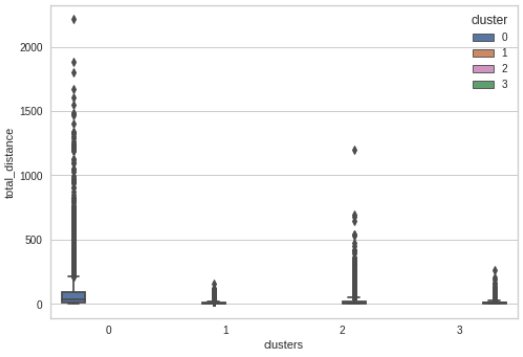
(a) Total nights of stay by cluster.



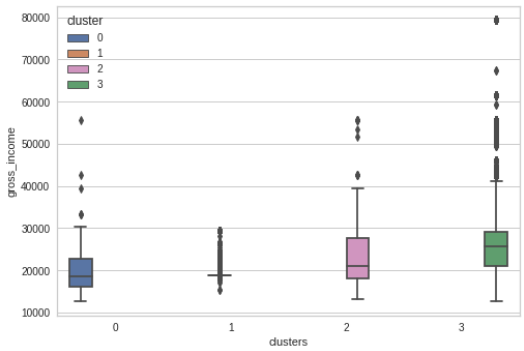
(b) Distance in Km. to the area by cluster.



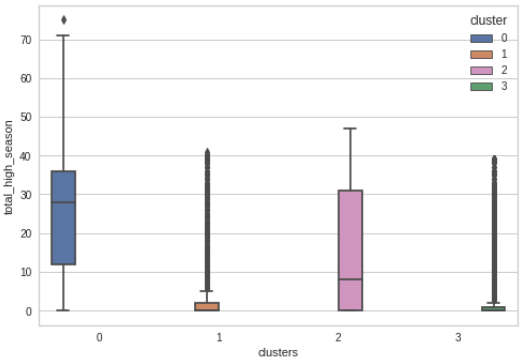
(c) Total entries to the area by cluster.



(d) Distance run in area by cluster.



(e) Average gross income by cluster.



(f) Total high season by cluster.

FIGURE 5.14: Box plots for ℓ^2 normalization.

Chapter 6

Discussions

Table 6.1 shows the equivalence by clusters and percentage of the total set for the two normalizations analyzed. For the group of registered residents, we can see that both normalization methods group them into a single cluster (cluster 3 in min-max and 0 in ℓ^2). However, there is a 2.62% difference in the size of these clusters, with the ℓ^2 cluster size being smaller. The min-max normalization distinguishes between foreign visitors and foreign non-registered residents (clusters 1 and 6, respectively), while the ℓ^2 normalization groups all foreign individuals into a single cluster (cluster 1). The clusters of national non-registered residents are also similar in both normalization methods (cluster 4 in min-max and 2 in ℓ^2). Still, there is a 4.19% difference in the size of these clusters, with the size of the ℓ^2 cluster also being smaller. Finally, the ℓ^2 normalization groups all national visitors into a single cluster (cluster 3), while the min-max normalization divides these into three distinct clusters (clusters 0, 2 and 5). It should be noted that in the ℓ^2 normalization, cluster 3 is larger than the sum of clusters 0, 2 and 5, because it contains individuals with resident behaviors that were not included in the other clusters. This explains the significant differences in the sample sizes of clusters 0 and 2 compared to their equivalents in the min-max normalization.

The min-max normalization seems more efficient since it allows a more detailed segmentation of individuals than ℓ^2, and ℓ^2 shows more outliers in the box plots for all the variables. While min-max seems to distinguish the residents from the visitors, with the variable representing the number of nights spent in the area, ℓ^2 seems to have a clear segmentation on the distance to their home. Hence, for our purposes, min-max offers better segmentations. In addition, min-max detects atypical behaviors of individuals not officially registered as residents of the area, but that behave as residents. In contrast, the ℓ^2 normalization could be useful for excluding foreigners from the analysis and focusing only on comparing registered and non-registered residents at the national level, grouping visitors in a single cluster.

Normalization								
Min-max	Nº cluster	3	1	6	4	0	2	5
	% sample	11.17%	6.11%	3.05%	8.55%	15.04%	8.58%	47.50%
ℓ^2	Nº cluster	0	1		2	3		
	% sample	8.55%	8.76%		4.36%	78.33%		

TABLE 6.1: Equivalence of the clusters made for each normalization.

In summary, the work presents an effective pipeline for clustering analysis, using data from different sensors and sources to detect registered and non-registered residents, and visitors; and their behavior in a given area. We have selected an optimal clustering algorithm based on the data distribution and two potential normalization algorithms. We found that the min-max normalization was the most effective for detailed segmentation of individuals and their visiting behavior in the area, and detection of atypical behavior of individuals not registered as residents of the area, but showing resident behavior. The ℓ^2 normalization could be useful in specific situations requiring a distinction from the region of origin. The information obtained from this analysis can help area managers to create personalized strategies for retaining specific tourists based on income or provenance and encouraging overnight stays, thereby generating wealth in the area and reducing the number of vehicles moving inside the area with other policies. This approach has important tourism planning and sustainability implications and is extensible to different regions. Our pipeline and analysis could also assist data analysts in improving their solutions and making informed decisions.

Chapter 7

Conclusions

In this study, we have proposed a pipeline to merge data from different sources in smart villages and analyzed these data to infer the traffic behavior in the area, providing policymakers the first step to understanding their traffic, which could lead to implementing effective policies aim towards a sustainable tourism. By doing so, we have accomplished the objectives outlined in Section 1.2 regarding the investigation and analysis of traffic behavior in smart villages. The following objectives were successfully attained:

- We analyzed the suitability of specific algorithms for different data distributions, enabling us to make informed decisions during the analysis process (see in Chapter 3).
- We extensively reviewed previous works on clustering applied to mobility patterns, information fusion, and traffic management, helping us leverage existing knowledge and identify potential gaps in the current research landscape (see in Chapter 2).
- We discussed the design and deployment process of the infrastructure used to collect data on vehicle movements in smart villages (see in Section 3.7).
- We analyzed the collected data and performed statistical analyses to identify patterns and correlations among variables from different data sources (see in Chapter 5).
- We evaluated various normalization techniques to preprocess the collected data, considering the effectiveness of each technique in improving result quality (see in Chapter 5).
- We applied the studied clustering algorithms to the collected data and evaluated the visualization and explainability of the results, aiming to select the most suitable algorithms for our analysis (see in Chapter 5).
- We assessed the performance of the proposed clustering pipeline and the selected normalization methods. We discussed the different segmentations obtained and the usability of each studied method (see in Chapter 6).

By achieving these objectives, our work contributes to understanding traffic behavior in smart villages. The information gained from this research can help improve

traffic flow, enhance visitor experience, and optimize resource allocation in these regions. Our findings serve as a valuable resource for the design and implementation of effective traffic and tourism management systems and policies in smart villages.

7.1 Limitations

One of the main limitations of this study lies in the inability to validate the unregistered residents identified through the segmentation of the problem. These unregistered residents are identified through mobility patterns extracted from available data sources. However, due to the lack of real information about them, their participation in such patterns cannot be verified. The lack of labels or real information about the data points is a common limitation in unsupervised learning algorithms. These algorithms are based on finding patterns and structures in the data without the guidance of predefined labels. However, when applying these methodologies in practical situations, it is important to recognize this limitation and consider it when interpreting the results and the practical implications of the study. There is a need for further research and ways to obtain more complete data, e.g., through questionnaires, that would allow an accurate understanding of mobility patterns in areas where the unregistered population may play a significant role.

Another important limitation of this study is the use of only three principal components (PCA) to represent the data in the analysis of mobility patterns. Cluster visualization is a crucial aspect in the analysis performed, but it is clear that we cannot visualize more than three dimensions using conventional plots. When considering Andrew Curves as an alternative to visualize more than three components, we face another limitation. These curves allow us to represent multiple variables and their relationships by plotting each observation as a curve on a two-dimensional graph. However, in problems with a large amount of data, it is difficult to visually interpret Andrew curves due to the overlapping and increasing visual complexity as more variables are added. This makes it difficult to identify and understand more complex and subtle mobility patterns that might be present in the data. Given this limitation, it is important to recognize that the representation and visualization of mobility patterns in this study are restricted by the reduced dimensionality and visualization techniques used. For a more detailed understanding of mobility patterns, it would be advisable to explore more advanced dimensionality reduction and visualization techniques that can address these challenges and provide a more complete and accurate representation of mobility patterns in future research.

7.2 Future Work

As future work, the project itself has offered us interesting opportunities to enrich and expand the research conducted in this work. One possible direction is to cross-reference the data used in this study with information generated by additional sensors (waste and motion sensors to detect persons in local businesses and streets), which have also been implemented in the project, and have the potential to provide valuable information that would complement existing vehicle data.

On the one hand, the incorporation of waste sensors would allow the analysis of waste generation and management patterns in relation to the identified mobility patterns. By correlating the amount and type of waste generated with traffic activity

in a specific area, deeper insights into the environmental impact and sustainable waste management needs in the context of urban mobility could be gained. This cross-referenced information could be useful to develop more efficient collection and recycling strategies.

On the other hand, the use of motion sensors that collect information on incoming and outgoing people counts and gauging could provide additional information on the mobility patterns of residents and visitors. By analyzing the number of people present at certain locations at different times, a more detailed understanding of people flows and their relationship to vehicle traffic could be obtained. This would allow the identification of time intervals, where pedestrian and vehicular mobility patterns are intertwined. These findings would be valuable for planning and designing urban infrastructures that promote sustainable mobility and transportation efficiency, taking into account both pedestrians and drivers.

Future work in this field could gain significant benefits from integrating data from waste and motion sensors. This integration would allow a more complete understanding of mobility patterns in smart villages by providing relevant information on environmental aspects, waste management and human behavior. It would be desirable to explore techniques and methodologies to combine and analyze these additional data in order to obtain a comprehensive view of urban mobility and its strategic implications. In addition, an integration of these indicators into a strategic dashboard could be carried out in order to merge the information and facilitate decision-making by local managers.

Bibliography

[1] Mihael Ankerst et al. "OPTICS: Ordering points to identify the clustering structure". In: *ACM Sigmod record* 28.2 (1999), pp. 49–60.

[2] Luigi Atzori, Antonio Iera, and Giacomo Morabito. "The internet of things: A survey". In: *Computer networks* 54.15 (2010), pp. 2787–2805.

[3] Mohammed Ayub and El-Sayed M El-Alfy. "Impact of normalization on BiLSTM based models for energy disaggregation". In: *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*. IEEE. 2020, pp. 1–6.

[4] Xiwen Bai et al. "A Data-Driven Iterative Multi-Attribute Clustering Algorithm and its Application in Port Congestion Estimation". In: *Available at SSRN 4086627* (2022).

[5] Jean-Patrick Baudry. "CLADAG 2015. Book of Abstracts". In: ISBN: 978888467749-9. 2015. Chap. Estimation and model selection for model-based clustering with the conditional classification likelihood. ISBN: 978888467749-9.

[6] Javier Béjar Alonso. *K-means vs Mini Batch K-means: a comparison*. Tech. rep. Universidad Politecnica de Madrid, 2013.

[7] Asma Belhadi et al. "Deep learning for pedestrian collective behavior analysis in smart cities: A model of group trajectory outlier detection". In: *Information Fusion* 65 (2021), pp. 13–20.

[8] Kamal Berahmand et al. "A novel method of spectral clustering in attributed networks by constructing parameter-free affinity matrix". In: *Cluster Computing* 25.2 (2022), pp. 869–888.

[9] Maria Bermudez-Edo, Payam Barnaghi, and Klaus Moessner. "Analysing real world data streams with spatio-temporal correlations: Entropy vs. Pearson correlation". In: *Automation in Construction* 88 (2018), pp. 87–100.

[10] A Bertuglia et al. "Reverse migration: from the city to the countryside. The Spanish case of Alpujarra Granadina." In: *Agriregionieuropa* 7.27 (2011), pp. 62–64.

[11] Macià Blázquez-Salom, Magdalena Cladera, and Maria Sard. "Identifying the sustainability indicators of overtourism and undertourism in Majorca". In: *Journal of Sustainable Tourism* (2021), pp. 1–25.

[12] Daniel Bolaños-Martinez, Maria Bermudez-Edo, and Jose Luis Garrido. "Clustering Study of Vehicle Behaviors Using License Plate Recognition". In: *Proceedings of the International Conference on Ubiquitous Computing & Ambient Intelligence (UCAmI 2022)*. Springer. 2022, pp. 784–795.

[13] Tadeusz Caliński and Jerzy Harabasz. "A dendrite method for cluster analysis". In: *Communications in Statistics-theory and Methods* 3.1 (1974), pp. 1–27.

[14] Oded Cats and Francesco Ferranti. "Unravelling individual mobility temporal patterns using longitudinal smart card data". In: *Research in Transportation Business & Management* 43 (2022), p. 100816.

[15] Roger Pueyo Centelles et al. "A lora-based communication system for coordinated response in an earthquake aftermath". In: *Multidisciplinary Digital Publishing Institute Proceedings* 31.1 (2019), p. 73.

[16] Dorin Comaniciu and Peter Meer. "Mean shift: A robust approach toward feature space analysis". In: *IEEE Transactions on pattern analysis and machine intelligence* 24.5 (2002), pp. 603–619.

[17] David L Davies and Donald W Bouldin. "A cluster separation measure". In: *IEEE transactions on pattern analysis and machine intelligence* 2 (1979), pp. 224–227.

[18] Zohreh Doborjeh et al. "Artificial intelligence: a systematic review of methods and applications in hospitality and tourism". In: *International Journal of Contemporary Hospitality Management* 34.3 (2022), pp. 1154–1176.

[19] Álvaro Daniel Rodríguez Escudero. "La Alpujarra granadina: un espacio rural diverso y complejo. De Sierra Nevada al litoral". In: *Nuevas realidades rurales en tiempos de crisis: territorios, actores, procesos y políticas: XIX Coloquio de Geografía Rural de la Asociación de Geógrafos Españoles y II Coloquio Internacional de Geografía Rural*. Universidad de Granada. 2018, pp. 782–794.

[20] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *International Conference on Knowledge Discovery and Data Mining (KDD)*. Vol. 96. 34. 1996, pp. 226–231.

[21] Anna Forster and Amy L Murphy. "CLIQUE: Role-free clustering with Q-learning for wireless sensor networks". In: *2009 29th IEEE International Conference on Distributed Computing Systems*. IEEE. 2009, pp. 441–449.

[22] Brendan J Frey and Delbert Dueck. "Clustering by passing messages between data points". In: *science* 315.5814 (2007), pp. 972–976.

[23] Jim Frost et al. "Multicollinearity in regression analysis: problems, detection, and solutions". In: *Statistics by Jim* (2017).

[24] Rafael Gallardo García et al. "Comparison of Clustering Algorithms in Text Clustering Tasks". In: *Computación y Sistemas* 24.2 (2020), pp. 429–437.

[25] Francisco M Garcia-Moreno et al. "A machine learning approach for semi-automatic assessment of IADL dependence in older adults with wearable sensors". In: *International Journal of Medical Informatics* 157 (2022), p. 104625.

[26] Aaron Gutiérrez et al. "Profiling tourists' use of public transport through smart travel card data". In: *Journal of Transport Geography* 88 (2020), p. 102820.

[27] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. "Clustering algorithms and validity measures". In: *Proceedings Thirteenth International Conference on Scientific and Statistical Database Management. SSDBM 2001*. IEEE. 2001, pp. 3–22.

[28] Graham Haughton and Colin Hunter. *Sustainable cities*. ISBN: 185302 234 9. Routledge, 2004. ISBN: 185302 234 9.

[29] Henderi Henderi, Tri Wahyuningsih, and Efana Rahwanto. "Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer". In: *International Journal of Informatics and Information Systems* 4.1 (2021), pp. 13–20.

[30] Zhengyu Hu. "Initializing the EM algorithm for data clustering and sub-population detection". PhD thesis. The Ohio State University, 2015.

[31] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[32] Enrique Navarro Jurado and Carmen Carvajal Gutiérrez. "Extranjeros jubilados: ¿residentes no empadronados o turistas residenciales? Metodología para la cuantificación de la población no empadronada". In: *BAETICA. Estudios de Historia Moderna y Contemporánea* 31 (2009), pp. 61–90.

[33] Elisabeth Kastenholz, Maria João Carneiro, and Celeste Eusébio. "Diverse socializing patterns in rural tourist experiences–a segmentation analysis". In: *Current Issues in Tourism* 21.4 (2018), pp. 401–421.

[34] Kyoungok Kim. "Spatial contiguity-constrained hierarchical clustering for traffic prediction in bike sharing systems". In: *IEEE Transactions on Intelligent Transportation Systems* 23.6 (2021), pp. 5754–5764.

[35] Billy Pik Lik Lau et al. "A survey of data fusion in smart city applications". In: *Information Fusion* 52 (2019), pp. 357–374.

[36] Huanhuan Li et al. "Unsupervised hierarchical methodology of maritime traffic pattern extraction for knowledge discovery". In: *Transportation Research Part C: Emerging Technologies* 143 (2022), p. 103856.

[37] Mengdan Lin and Xuelin Zhao. "Application research of neural network in vehicle target recognition and classification". In: *2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*. IEEE. 2019, pp. 5–8.

[38] Yanchi Liu et al. "Understanding of internal clustering validation measures". In: *2010 IEEE international conference on data mining*. IEEE. 2010, pp. 911–916.

[39] Manjarini Mallik, Ayan Kumar Panja, and Chandreyee Chowdhury. "Paving the way with machine learning for seamless indoor-outdoor positioning: A survey". In: *Information Fusion* (2023).

[40] Ana Jiménez Martín et al. "Affinity propagation clustering for older adults daily routine estimation". In: *2021 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE. 2021, pp. 1–7.

[41] Leland McInnes, John Healy, and Steve Astels. "hdbscan: Hierarchical density based clustering." In: *J. Open Source Softw.* 2.11 (2017), p. 205.

[42] Nargess Memarsadeghi et al. "A fast implementation of the ISODATA clustering algorithm". In: *International Journal of Computational Geometry & Applications* 17.01 (2007), pp. 71–103.

[43] Md Ashifuddin Mondal and Zeenat Rehena. "Identifying traffic congestion pattern using k-means clustering technique". In: *2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU)*. IEEE. 2019, pp. 1–5.

[44] Fagner Sutel de Moura and Christine Tessele Nodari. "Application of the Affinity Propagation Clustering Technique to obtain traffic accident clusters at macro, meso, and micro levels". In: *arXiv preprint arXiv:2202.05175* (2022).

[45] Stefan Neubig et al. "Data-driven Initiatives of Destinations Supporting Sustainable Tourism". In: (2022).

[46] Zhaolong Ning, Jun Huang, and Xiaojie Wang. "Vehicular fog computing: Enabling real-time traffic management for smart cities". In: *IEEE Wireless Communications* 26.1 (2019), pp. 87–93.

[47] Angel Paniagua. "Smart villages in depopulated areas". In: *Smart Village Technology: Concepts and Developments* (2020), pp. 399–409.

[48] Jinwan Park, Jungsik Jeong, and Youngsoo Park. "Ship trajectory prediction based on bi-LSTM using spectral-clustered AIS data". In: *Journal of marine science and engineering* 9.9 (2021), p. 1037.

[49] Subbulakshmi Pasupathi et al. "Trend analysis using agglomerative hierarchical clustering approach for time series big data". In: *The Journal of Supercomputing* 77 (2021), pp. 6505–6524.

[50] SGOPAL Patro and Kishore Kumar Sahu. "Normalization: A preprocessing stage". In: *arXiv preprint arXiv:1503.06462* (2015).

[51] Harald Pechlaner, Elisa Innerhofer, and Greta Erschbamer. *Overtourism: Tourism management and solutions*. Routledge, 2019.

[52] Maycon Leone Maciel Peixoto et al. "A traffic data clustering framework based on fog computing for VANETs". In: *Vehicular Communications* 31 (2021), p. 100370.

[53] Vicente Pinilla, María-Isabel Ayuda, and Luis-Antonio Sáez. "Rural depopulation and the migration turnaround in Mediterranean Western Europe: a case study of Aragon". In: *Journal of Rural and Community Development* 3.1 (2008).

[54] Kemal Polat and Umit Sentürk. "A Novel ML Approach to Prediction of Breast Cancer: Combining of mad normalization, KMC based feature weighting and AdaBoostM1 classifier". In: *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. Ieee. 2018, pp. 1–4.

[55] Bagus Priambodo, Azlina Ahmad, and Rabiah Abdul Kadir. "Predicting traffic flow propagation based on congestion at neighbouring roads using hidden Markov model". In: *IEEE Access* 9 (2021), pp. 85933–85946.

[56] Douglas Reynolds. "Gaussian Mixture Models". In: *Encyclopedia of Biometrics*. Ed. by Stan Z. Li and Anil Jain. Boston, MA: Springer US, 2009, pp. 659–663. ISBN: 978-0-387-73003-5.

[57] Joaquín Amat Rodrigo. *Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE*. Accessed: 2023-3-29, available under a Attribution 4.0 International (CC BY 4.0). 2017.

[58] Vicente Rodriguez, Gloria Fernandez-Mayoralas, and Fermina Rojo. "International retirement migration: Retired Europeans living on the Costa del Sol, Spain". In: *Population Review* 43.1 (2004), pp. 1–36.

[59] Peter J Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.

[60] Fernando Terroso Sáenz, Francisco Arcas-Tunez, and Andres Muñoz. "Nationwide touristic flow prediction with Graph Neural Networks and heterogeneous open data". In: *Information Fusion* 91 (2023), pp. 582–597.

[61] Ahmed M Serdah and Wesam M Ashour. "Clustering large-scale data based on modified affinity propagation algorithm". In: *Journal of Artificial Intelligence and Soft Computing Research* 6.1 (2016), pp. 23–33.

[62] Gholamhosein Sheikholeslami, Surojit Chatterjee, and Aidong Zhang. "Wavecluster: A multi-resolution clustering approach for very large spatial databases". In: *VLDB*. Vol. 98. 1998, pp. 428–439.

[63] Hiroaki Shiokawa. "Scalable affinity propagation for massive datasets". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 11. 2021, pp. 9639–9646.

[64] Steffen Staab et al. "Intelligent systems for tourism". In: *IEEE intelligent systems* 17.6 (2002), pp. 53–66.

[65] Haodong Sun et al. "Identifying tourists and locals by K-means clustering method from mobile phone signaling data". In: *Journal of Transportation Engineering, Part A: Systems* 147.10 (2021), p. 04021070.

[66] Ulrike Von Luxburg. "A tutorial on spectral clustering". In: *Statistics and computing* 17.4 (2007), pp. 395–416.

[67] Wei Wang, Jiong Yang, Richard Muntz, et al. "STING: A statistical information grid approach to spatial data mining". In: *VLDB*. Vol. 97. Citeseer. 1997, pp. 186–195.

[68] Zijia Wang et al. "Identifying urban functional areas and their dynamic changes in Beijing: using multiyear transit smart card data". In: *Journal of Urban Planning and Development* 147.2 (2021), p. 04021002.

[69] Stephan D Whitaker. "Did the COVID-19 pandemic cause an urban exodus?" In: *Cleveland Fed District Data Brief* 20210205 (2021).

[70] Allan M Williams, Russell King, and Tony Warnes. "A place in the sun: International retirement migration from northern to southern Europe". In: *European urban and regional studies* 4.2 (1997), pp. 115–134.

[71] Dan Xie and Yu He. "Marketing Strategy of Rural Tourism Based on Big Data and Artificial Intelligence". In: *Mobile Information Systems* 2022 (2022).

[72] Miin-Shen Yang, Chien-Yo Lai, and Chih-Ying Lin. "A robust EM clustering algorithm for Gaussian mixture models". In: *Pattern Recognition* 45.11 (2012), pp. 3950–3961.

[73] Wenbin Yao et al. "Analysis of Key Commuting Routes Based on Spatiotemporal Trip Chain". In: *Journal of Advanced Transportation* 2022 (2022).

[74] Wenbin Yao et al. "Understanding travel behavior adjustment under COVID-19". In: *Communications in Transportation Research* (2022), p. 100068.

[75] Wenbin Yao et al. "Understanding vehicles commuting pattern based on license plate recognition data". In: *Transportation Research Part C: Emerging Technologies* 128 (2021), p. 103142.

[76] Bin YU and Jun XIONG. "A Novel WSN Traffic Anomaly Detection Scheme Based on BIRCH". In: *Journal of Electronics & Information Technology* 44.1 (2022), pp. 305–313.

[77] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. "BIRCH: an efficient data clustering method for very large databases". In: *ACM sigmod record* 25.2 (1996), pp. 103–114.

[78] Shengjie Zhao et al. "Hyper-clustering enhanced spatio-temporal deep learning for traffic and demand prediction in bike-sharing systems". In: *Information Sciences* 612 (2022), pp. 626–637.