FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Understanding urban mobility patterns of tourists through data extraction

João Manuel Angélico Gonçalves

DISSERTATION

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Marta Campos Ferreira

Co-Supervisor: Teresa Galvão Dias

July 31, 2021

Understanding urban mobility patterns of tourists through data extraction

João Manuel Angélico Gonçalves

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: Prof. Rosaldo Rossetti
External Examiner: Prof. Marco Veloso
Supervisor: Prof. Marta Campos Ferreira
Co-Supervisor: Prof. Teresa Galvão Dias

July 31, 2021

Abstract

Electronic fare payment systems have gained much popularity around the world. These systems adopt a convenient and almost instantaneous payment process for public transport, making it a more comfortable process for the passengers while also gathering accurate and detailed data regarding on-board transactions in public transport. Much information can be extracted about public transport passengers, such as travel patterns, activities performed, and travel behavior. Despite the continuous growth of studies regarding these systems, there is still a lack of research to understand occasional passengers' movement, such as tourists. By doing so, the knowledge extracted would allow the definition of mobility services adapted to tourists' needs, which would improve the tourist's experience in the city and, consequently, increase the number of visitors for the city and its points-of-interest. This work explores AFC data to understand the mobility profiles of tourists. It relies on data from the AFC system of Porto, Portugal, from the year 2019. The ticketing system implemented in the Metropolitan Area of Porto offers different types of tickets in order to meet the different needs of citizens. For example, tour tickets are mostly used by tourists and monthly subscriptions by regular users. Thus, to understand tourists' mobility profiles, we developed a scoring system capable of identifying tourists in an AFC ticketing system. This model was applied to data from the AFC system in Porto, allowing the identification of tourists among the data. After having the data precisely separated, the AFC data was then submitted to data mining techniques to capture tourists' mobility patterns. These techniques focus on clustering-based approaches and allow the definition of spatial patterns, temporal patterns, and spatial-temporal patterns, a combination of the previous two. Firstly, the spatial patterns allowed us to indicate the most visited areas by tourists. Secondly, with the temporal patterns, we were able to identify the busiest hours, days of the week and seasons for non-frequent passengers. Finally, it was possible to profile public transport usage for both space and time with spatial-temporal patterns. At last, the work dedicates a section to present and discuss the obtained results and, for better comprehension, these results are also compared with data obtained from the city's points of interest, based on Google Places API. The overall analysis allowed us to verify that our scoring method was accurate while separating tourists from regular users, though it faced some difficulties while identifying tourists from occasional users. The analysis also allowed us to characterize tourists' spatial and temporal dynamics within the city and identify different mobility patterns and the city's areas most frequented by tourists. By crossing the two datasets, it was also possible to understand the services that are located closer to the stations most used by tourists, allowing to characterize the type of most visited places. This study represents an advance in the literature and opens doors to the definition of policies to promote less visited places and mobility services adapted to tourists' needs, resulting in a positive impact on the city's economy and the overall enjoyment of the city for tourists.

Keywords: Public Transports, Smart card data, Tourism Detection, Travel Behaviour

Acknowledgements

First and foremost, I am highly grateful to my supervisors, Marta Campos Ferreira and Teresa Galvão Dias, for their guidance, constant support, and patience during this dissertation.

I would like to extend my sincere thanks to my friends and relatives who always comforted me and made this journey enjoyable.

Finally, I would like to express my gratitude to my parents. Without their tremendous understanding, encouragement, and emotional support, it would have been impossible for me to complete this study.

João Manuel Angélico Gonçalves

Contents

List of Figures

List of Tables

Abbreviations

AFC	Automated Fare Collection
CRISP-DM	Cross Industry Standard Process for Data Mining
CP	Comboios de Portugal
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
LPA	Latent Profile Analysis
MAP	Metropolitan area of Porto
OD	Origin-Destination
POI	Point of Interest
POIs	Points of Interest
RUP	Regular Users' Percentage
STCP	Sociedade de Transportes Colectivos do Porto
TIP	Transportes Intermodais do Porto
TP	Tourists' Percentage

Chapter 1

Introduction

1.1 Problem Context

With the rapid advancements of technology over the years, it was an important step for public transportation networks to adopt an automated ticketing system. And with that in mind, in 1990, there was an emergence of Automated Fare Collection (AFC) systems around the world. These systems provide an easy, convenient and instantaneous payment process through the use of smart travel cards while also gathering detailed data for each trip. Even though Automated Fare Collection (AFC) systems' main goal was to automate and manage the ticketing process for fare collection, the detailed and vast amount of data collected regarding onboarding transactions presented great opportunities in extracting valuable knowledge about the public transport passengers.

Due to this knowledge's global importance, a lot of information and conclusions have been extracted in the past regarding frequent passengers of public transports. However, there has not been much knowledge gained from the occasional passengers, such as tourists.

The Metropolitan Area of Porto (MAP) public transport network is composed of private and public buses, light rail, and trains and adopted an Automated Fare Collection (AFC) system in 2002 called Andante. Andante is an entry-only AFC system with a distance-based fare structure that covers the metropolitan area of Porto. The fare media of Andante are contactless travel cards that can be used on buses from several operators, as well as on metro and railways.

For this dissertation, we explored the data from the AFC system of the city of Porto, Portugal, from the year 2019 to discover mobility patterns of tourists within the city.

1.2 Motivation

Identical to the demographic rise in urban centers, there has been a massive increase in tourist activity. During the last 30 years, international tourist arrivals have more than tripled, and the importance of this area in various countries' economies today transcends 10% of their gross domestic product [9].

Although this growth in tourism and its economic impact worldwide is generally positive, it is placing tough challenges on many cities that were not initially prepared to receive such an amount of tourists. These challenges range from tourism gentrification to mobility concerns such as the overuse of road networks and public transports [2].

Nowadays, more than ever, it is essential to understand tourists' mobility patterns to make tourism more sustainable and help in the definition of tourism-oriented policies.

Since this growth is also evident in Porto's city (shown in Figure 1.1), our case study, the AFC data from this city, can be utilized to explore and study this theme.

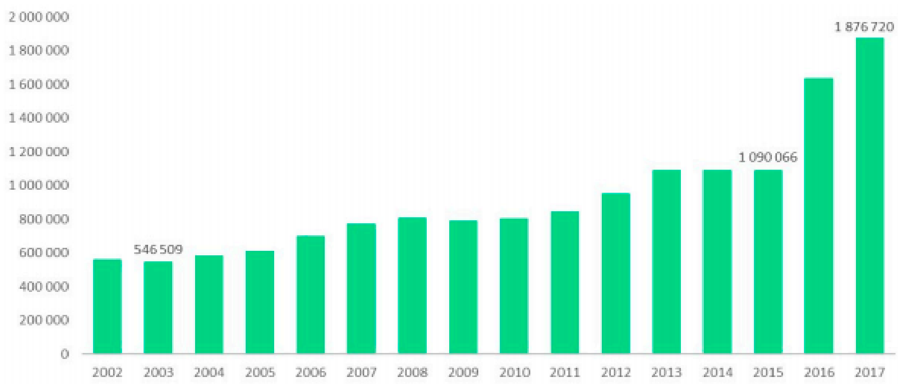


Figure 1.1: Number of guests in hotel establishments in the municipality of Porto [14]

1.3 Objectives

This dissertation aims to extract knowledge from AFC data from the city of Porto to identify tourists and then to understand their mobility profiles. From this knowledge, we will be able to formalize spatial patterns, temporal patterns, and spatial-temporal patterns. Firstly, the spatial patterns will allow us to indicate the most visited areas by tourists and to identify station types. Secondly, with the temporal patterns, we will identify the busiest hours, days of the week and seasons for tourist passengers. Finally, with the spatial-temporal patterns, it will be possible to profile the stations' public transport temporal usage.

In addition to this, it will also be important to contextualize this information by complementing it with an analysis of a second dataset consisting of the city of Porto's points-of-interest, based on Google Places API. By crossing the two datasets, it will be possible to understand the services that are located closer to the stations most used by tourists, allowing to characterize the type of most visited places.

1.4 Document Structure

In addition to the introduction, this dissertation is organized in five more chapters. Chapter 2 will consist on a literature review of Data Mining and Automated Fare Collection Systems. In chapter 3 we characterize our problem and present the methodology followed for our solution. In Chapter 4 we will demonstrate and explain our scoring system capable of identifying tourists through their validations. Chapter 5 will show all patterns created from the tourists' AFC data. Finally, in chapter 6, we present our conclusions and future work to be done.

Chapter 2

Literature Review

This chapter provides an introduction to the literature related to the use of data mining techniques in Automated Fare Collection Systems to extract passenger patterns.

Firstly, we provide an overview of Data Mining Techniques with a focus on Cross-Industry Standard Process for Data Mining (CRISP-DM) since it is the methodology that will be applied in this project. The second section provides an overview of the Automated Fare Collection Systems, the Alighting point estimation, the use of AFC data for Travel Behavior and Travel Pattern Analysis, and how the AFC data was filtered for tourism analysis.

2.1 Data Mining

For this project, we will use data mining techniques to extract knowledge from the AFC data provided to us.

Data Mining can be defined as a process capable of extracting knowledge and patterns from data sets in order to predict an outcome. Different methodologies can be used in the construction of a data mining model, as Figure 2.1 shows, but the methodology that will be used while developing our project is the Cross-Industry Standard Process for Data Mining(CRISP-DM).

2.1.1 CRISP-DM

CRISP-DM is a methodology that provides an overview of a data mining project's life cycle and can be broken down into 6 different phases, as shown from the Figure 2.2. The different stages can be reversible if the results achieved are not ideal. The end of the process will only occur at the last stage of the cycle unless the project is terminated in an intermediate step. The 6 different phases of this methodology are Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. These are explained in detail below.

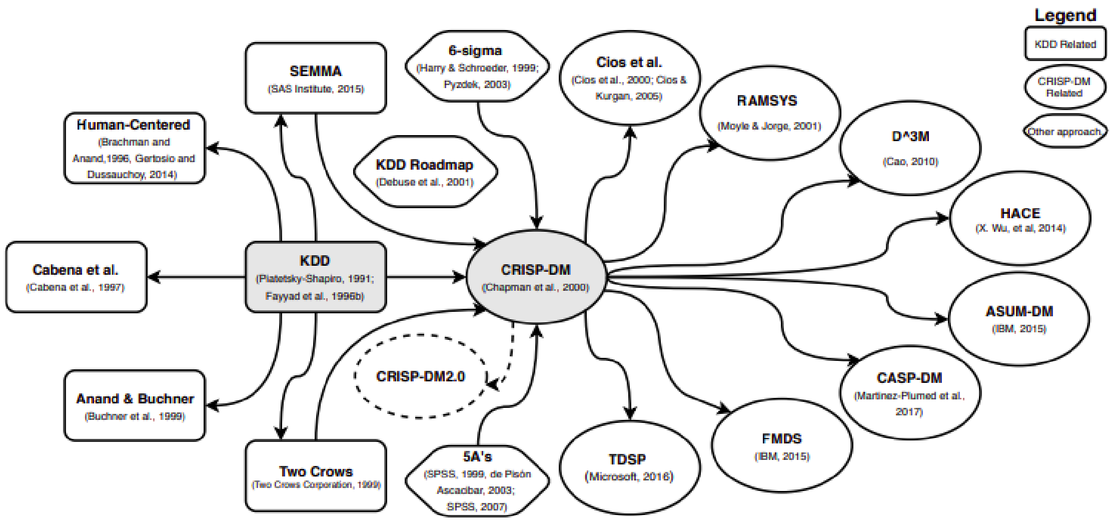


Figure 2.1: Data Mining Models and Methodologies

2.1.1.1 Business Understanding

This initial phase concentrates on understanding the project objectives and requirements from a business perspective in order to convert this information into a data mining problem to achieve the objectives. For this phase, the following tasks were identified:

- Determine the Business Goals;
- Assess the Situation;
- Determine the Data Mining Objectives;
- Produce the Plan for the Project;

2.1.1.2 Data Understanding

Firstly, the Data Understanding phase starts with a data collection and familiarization so we can later verify if the data is appropriate for our needs. For this phase, the following tasks were identified:

- Collect the Initial Data;
- Describe the Data;
- Explore the Data;
- Verify the Quality of the Data;

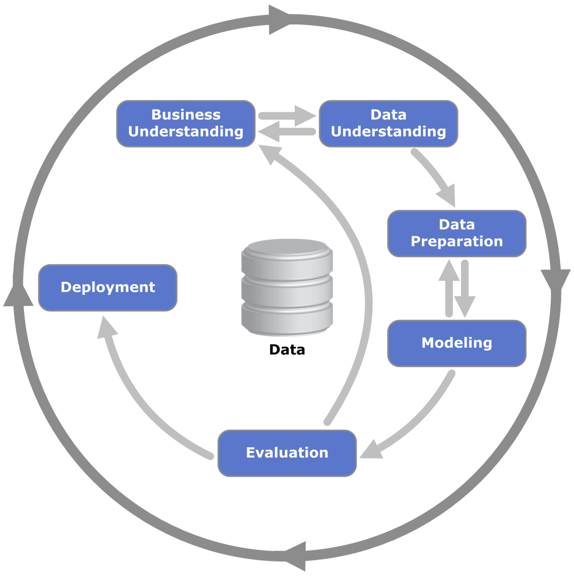


Figure 2.2: CRISP-DM Process

2.1.1.3 Data Preparation

In the Data Preparation phase, all refinements are done to the original raw data before it is ready to use for modeling. For the Data Preparation phase , the following five five tasks were identified:

- Select the data;
- Clean the data;
- Construct the data;
- Integrate the data;
- Format the data;

2.1.1.4 Modeling

In the modeling phase, techniques are applied to the data to search for useful patterns. This phase consists of the following tasks:

- Select the Modeling Technique;
- Generate the Test Design;
- Build and Assess the Model;

There is a wide variety of modeling techniques that can be applied in this phase, with some of them being anomaly detection, clustering, classification, association rule learning, and regression but, for this project, the most relevant data mining method will be **clustering**.

Clustering is a machine learning method that identifies similar data points in the dataset in order to organize a set of objects in subgroups. The cluster population should ideally be as heterogeneous as possible, but there should be homogeneity between clusters.

2.1.1.5 Evaluation

At the evaluation stage, it will be calculated the degree to which the model meets the business objectives previously defined. The Evaluation phase includes three tasks, with those being:

- Evaluate the results;
- Review the process;
- Determine the next steps;

2.1.1.6 Deployment

At the Deployment phase, the knowledge gathered during the modeling phase, will be organized and presented in a way that it can be put to practice. For the deployment phase, the following tasks were identified:

- Plan the deployment;
- Plan monitoring and maintenance;
- Report and Review the final results;

2.2 Automated Fare Collection Systems

Recently, as Automated Fare Collection Systems continue to be implemented worldwide, the number of research studies involving them has been increasing. These systems were introduced to replace the traditional tickets with smart cards that improve the payment mechanism and information flow while ensuring smoother passenger boarding. Blythe [6] gives a more precise explanation of how the public transport ticketing process was improved through smart cards.

There are several popular implementations of AFC systems, such as the Oyster card in London, the Octopus card in Hong Kong, the Navigo in Paris, the Bip! in Santiago, the OV-Chip in the Netherlands, and, of course, the Andante in Porto which will be the subject for our study.

Each fare collection system has its personal data collection approach, but what is constant in all of them, is that the information concerning journey transactions is automatically collected.

2.2.1 Alighting point estimation

Choosing an adequate strategy for the tickets' validation process is of the foremost importance in the implementation of an AFC system. This choice involves choosing between an entry-only system, an entry-exit system, or even a combination of both, depending on the transportation network.

Whereas in an entry-exit system, the validation is done at both origin and destination, for the entry-only system, the passenger will only validate his ticket at the origin station/bus stop. Some examples of smart cards and their systems are represented in the Table 2.1

Table 2.1: Examples of AFC Ticketing Systems

	Entry-Only	**Entry-Exit**
Oyster card	X	X
Octopus card	X	X
Navigator Card	X	
Navigo	X	X
Bip!		X
OV-Chip	X	X
EZ-Link		X
T-Money	X	
Andante	X	

There is a major concern for fare collection systems using an entry-only system, since each trip's destination is unknown, which is the case for the Andante AFC system. So, in order to overcome this issue, several studies explored ways of estimating the alighting point.

Barry et al. [5] used AFC data from the New York City subway system to present a method to estimate the destination point for each trip. To do such, it followed the assumptions that :

- After a trip, passengers will return to the destination of their previous trip station;
- At the end of a day, users will return to the station where they boarded for the first trip of that same day;

Later, Zhao et al. [27] while estimating the alighting point for the Chicago CTA system, added to the assumptions of Barry et al. [5] that:

- The maximum walking distance is 400 m;

By doing such, Zhao et al. could assign the nearest station to the next boarding bus stop within a 400 m radius as the alighting station.

Ali et al. [3] extracted data from the T-Money AFC system in order to analyze travel behavior. Since it his a entry-only system for some networks, it was necessary to determine the destination points of trips. To do such, in addition to using the assumptions mentioned by Barry et al. [5] and Zhao et al. [27], Ali also added that, for each trip:

- If the time difference between previous trip's alighting and current trips boarding is greater than 30 minutes, it is inferred as an activity, otherwise it is considered as a transfer point.

Once the trips are separated from trip segments, the chains are identified, and unnecessary data can be removed from the database.

Munizaga et al. [19] reconstructed the trip chain of users behind bip! cards by estimating the destination points from the information available. To do such, it incorporated additional constraints to the ones mentioned by Barry et al. [5] and Zhao et al. [27] such as looking at the position and time of the next boarding and assuming that a trip destination is close to the point where the first trip of the day began, completing a full cycle. By doing such, a success rate of 80% was achieved.

2.2.2 Travel Behavior and Travel Pattern Analysis

In this subsection, we will discuss various approaches that previous researchers adopted to characterize public transport passengers' travel behavior and patterns with the usage of clustering in AFC data.

Ortega-Tong [23] using a sample of AFC data generated from Oyster Cards, performed a classification of London's public transport passengers by using the K-Medoids clustering algorithm. Due to the vast amount of attributes stored for each validation, it was possible to perform a temporal-spatial analysis for the passengers' trip entry station and boardings with variables related to:

- Temporal variability, such as the travel frequency and journey start time;
- Spatial variability, such as the origin station regularity and the travel distance;
- Activity pattern variability.

Spatial patterns were also studied by analyzing the most commonly visited stations and comparing them with the entire population. From this, it was possible to verify that regular users' spatial travel patterns were actually similar to the whole population.

Finally, Ortega-Tong [23] directed his study for London's visitors by performing another temporal-spatial analysis. With that, he was able to draw valuable conclusions related to visitors, such as that most of their movement is in the center of the city and at off-peak hours and that their most common public transportation mode is the rail.

Namiot et al. [20] presented various data models to process data gathered from smart cards. With that, they were able to gain valuable information for transit patterns detection and trips discovery. While doing his research, Namiot et al. examined the paper [18].

Xiaolei et al. [18] developed a model for individual travel pattern recognition and travel regularity mining. The first step for this work consisted of generating the passengers' trip chains (series of trips made by a user daily). For this process, to associate several validations into a trip chain, the authors used a fixed temporal threshold. With that, it was possible to store, for each trip chain, the chain and card id, the date, the first boarding time, the last alighting time, the route sequence and the stop ID sequence.

Once the trip chain info has been composed, each transit rider's travel pattern is further examined by clustering the trip chains. To do such, they utilized the algorithm DBSCAN in the trip chains to detect travel patterns for passengers.

Finally, in order to detect regularity, Xiaolei et al. used the following features of the trip chain data for clustering:

- Number of travel days.

- Number of similar first boarding times since they represent the passenger temporal characteristics.

- Number of similar route sequences since they represent the spatial characteristics of the passenger.

- Number of similar stop ID sequences since they can contain detailed spatial similarity information.

Gutiérrez et al. [15], in order to detect public transport users' profiles used Latent Profile Analysis (LPA), a model-based clustering technique using the expectation-maximization algorithm.

The first analysis performed corresponded to clustering profiles from activity variables, such as the number of transactions, the card lifespan (enabled days), the number of days the card was used (active days) and the average number of consecutive transactions in any stop (avg group size). By clustering the AFC data with these variables, a five-profile solution was obtained with different values for the attributes used (shown in Figure 2.3).

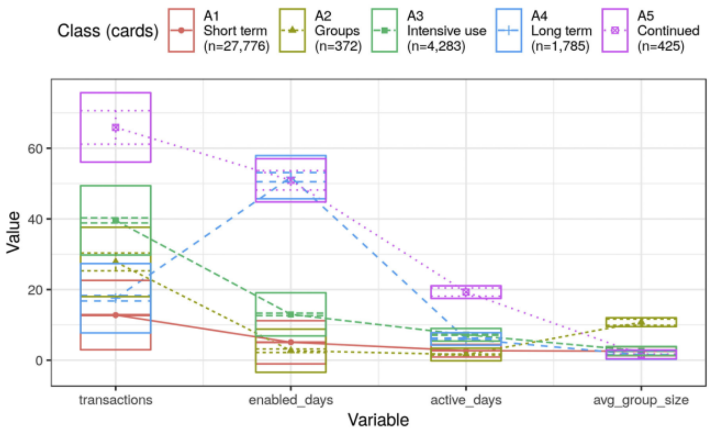


Figure 2.3: Descriptive statistics of the profiles obtained by means of LPA with activity variables

After that, Gutiérrez et al. clustered the riders' profiles from spatial variables, and the solution obtained consisted of 4 different profiles. From this, it was possible to analyze the spatial characteristics for each one of the clusters, and it was also possible to draw some conclusions on the concentration of transactions for certain areas, which is shown in Figure. 2.4.

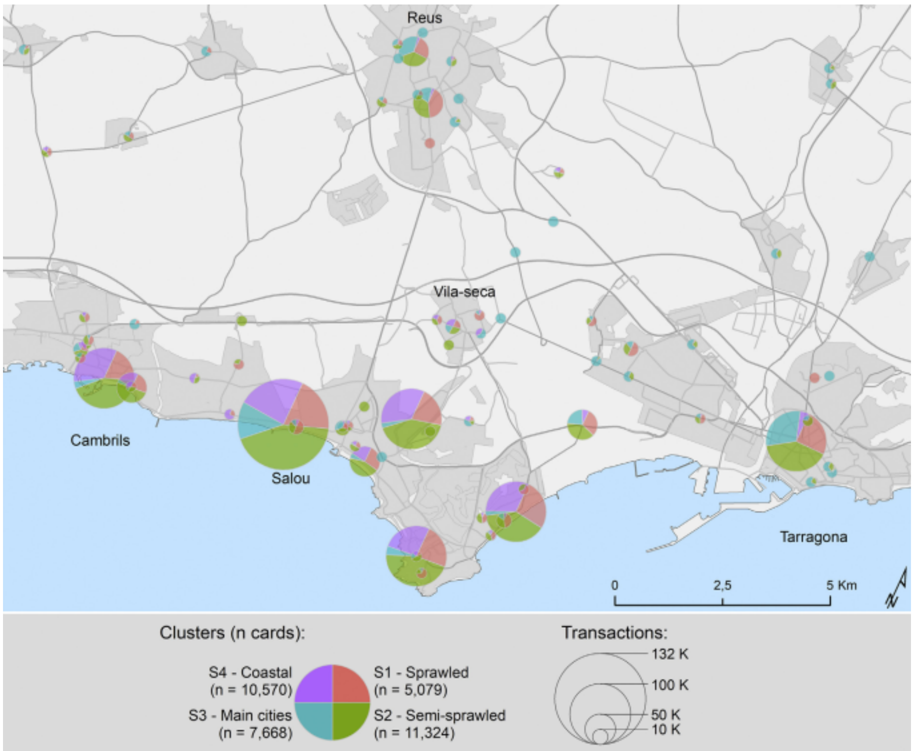


Figure 2.4: Transaction by bus stops and spatial activity profiles

As mentioned before, Agard et al. [1] developed a model for mining public transport user behavior from smart card data. The authors considered the whole database and performed a clustering technique to subdivide this data into several groups based on the trip patterns.

Due to the large amount of data collected, a first group was computed using a k-means method, which resulted in a 20 group partition. The result of the k-means clustering was then used as the input for another clustering algorithm but, this time, it was the Hierarchical Ascending Clustering method (HAC) that was applied. The processing of both these methods resulted in 4 clusters for the whole database.

With these four groups, Agard et al. [1] were able to perform a more in-depth analysis of their temporal attributes and draw conclusions about the busiest time zones and days for each cluster. It was also possible to verify how the variability changed over 12 weeks for passengers using the students' card.

Chakirov et al. [7] performed an analysis of AFC data from one day in the city-state of Singapore to demonstrate how the EZ-Link data can be used as an indicator of region travel behavior and to identify factors influencing peoples' decisions of traveling in public transports.

This analysis involves an analysis of temporal characteristics such as the temporal trip start distribution and identification of waiting times for each station. This made it possible to understand that the seat availability is valued for longer trips, so stations with a high volume of riders will sometimes hold higher waiting times.

Gan et al. [13] using data collected from smart cards in the Nanjing Metro System in China, managed to identify spatial-temporal ridership profiles for the metro stations while also exploring the relationship between these profiles and local Land-Cover and Land-Use (LCLU).

A cluster analysis was used to classify the Nanjing Metro stations according to their ridership patterns, using the k-means algorithm. The stations were aggregated into 7 clusters based on their hourly intervals for both boarding and alighting time.

The authors then extracted distinct spatial-temporal ridership profiles for each cluster to better understand the urban mobility patterns in a space and time perspective. With this, they were able to verify the busiest boarding and alighting hours in each station and cluster of stations, as is shown in Figure 2.5.

Finally, to contextualize the information extracted for each station, Gan et al. [13] related the clusters to the city's LCLU to understand heterogeneity in ridership patterns and highlight how points of interest motivate urban mobility in a daily cycle.

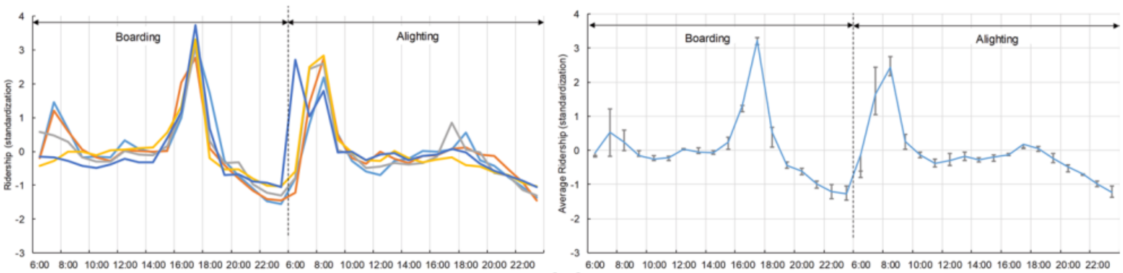


Figure 2.5: Daily ridership profile for cluster of employment-oriented stations

El Mahrsi, et al. [10] proposed two approaches of clustering smart card data from the metropolitan area of Rennes in France. The first one is a station-oriented operational point of view to determine how the stations' trips are distributed over time and, the second is passenger-focused to identify groups of passengers with similar boarding times aggregated into weekly profiles.

In order to identify the relationships between different stations, these were clustered according to the validations they receive from passengers. The method the authors used to cluster the stations was an adaptation of the Bike Sharing System station clustering method by Côme et al. [8] and it was followed with a Poisson mixture to build a model. After having the stations clustered then the results were studied based on their spatial-temporal activity characterizations and broken down into two categories, with them being:

- Balanced Stations - stable usage throughout the day with peaks during rush hours.
- Unbalanced Station - unbalanced usage throughout the day (transactions in a part of the day are very different to another part)

In the second phase of their study, El Mahrsi, et al. [10] clustered riders based on their temporal behavior to identify recurring patterns in the way passengers use public transportation. Since information about the user was necessary, only the data associated with regular users were considered.

The method that was used to cluster based on temporal behaviors relies on estimating a mixture of unigrams model [21] from the retained passengers' temporal profiles and was followed with an Expectation–maximization algorithm. With this step completed, the authors discovered 13 passenger clusters that were then characterized according to their spatial-temporal activity.

2.2.3 Tourism Analysis

Although there has been significant knowledge extracted from recurrent users, there is still a lack of research on reviewing the potentialities that AFC data could offer for public transport management in the tourism context.

To focus our research on non-frequent passengers, even though there has yet to been found a precise solution, it will be essential to separate accurately the data from tourists from the data of other passengers. Table 2.2 displays the methodology used by some examples of tourist estimation.

Table 2.2: Examples of Tourist Estimation Methods

Paper	Location	Method
[1]	Gatineau, Canada	Clustering
[15]	Tarragona, Spain	Card Type
[17]	-	Iterative Refinement
[23]	London, England	Card Type
[26]	Singapore	Iterative Refinement

Agard et al. [1] while mining public transport user behavior from smart card data, managed to divide the AFC data into four different clusters. By studying each cluster's activity, they were able to verify similar trip habits throughout the week and, with that, classify these groups of people as regular users, non-frequent users, students or elderlies.

Gutiérrez et al. [15] used data collected through smart travel cards in Tarragona (Spain) to profile the tourists within the city and, to disregard data from regular passengers, they focused their analysis on the T-10, a multi-personal transport card with no time limitations that tourists mainly use.

Ortega-Tong [23] like Gutiérrez et al., also opted to recognize a tourist by the ticket he uses. To learn London's visitors' behavior, he analyzed data from the Visitor Oyster Cards, a type of Oyster Card issued to tourists that can be purchased in advance and delivered to any country.

Xue et al. [26] using bus and train riding records from Singapore's EZ-Link cards, developed an approach to identify tourists from recurrent passengers. This approach consisted of 2 stages with those being:

1. **Station Ranking -** Score each station based on its attractiveness to tourists.

2. **Iterative Refinement -** Iteratively update the scores for each station and for the tourists who visited them in a graph. After the last iteration, the passenger can then be classified as a tourist or not, depending on its respective score.

Lu et al. [17] developed a framework that identifies tourists and subsequently conducts their preference analytics using city-scale public transportation data and, for its tourist identification system, Lu et al. opted to use the same approach as Xue et al. [26].

Chapter 3

Methodology

This chapter focuses on understanding the context and domain of the problem and explaining the methodological approaches we used while finding a solution to that problem.

3.1 Context

To better understand the problem in question, it is vital to analyze the domain on which it is inserted. For this reason, in this subsection, we perform an analysis of the Metropolitan Area of Porto Public Transport and the data provided from the AFC systems and the city's Points of Interest.

3.1.1 Metropolitan Area of Porto

Located on the North Coast of Portugal, the Metropolitan Area of Porto (MAP) covers 17 municipalities with an estimated area of 2.040 km2 that serves approximately 1,7 million inhabitants. This area is displayed in Figure 3.1.

The public transportation network mobility is composed of several transportation companies, with those being:

- **Metro do Porto** - Manages the Railway network
- **Sociedade de Transportes Colectivos do Porto (STCP)** - Manages the public bus network
- **Comboios de Portugal (CP)** - Manages the train network
- **29 Private bus companies**.

The Metropolitan Area of Porto's electronic ticketing system is an open (ungated) system that contains ticket readers at each railway/train station and each bus vehicle.

The fare media in MAP are contactless travel cards, named Andante, that can be utilized throughout the transportation network. The Andante system was created and is currently being

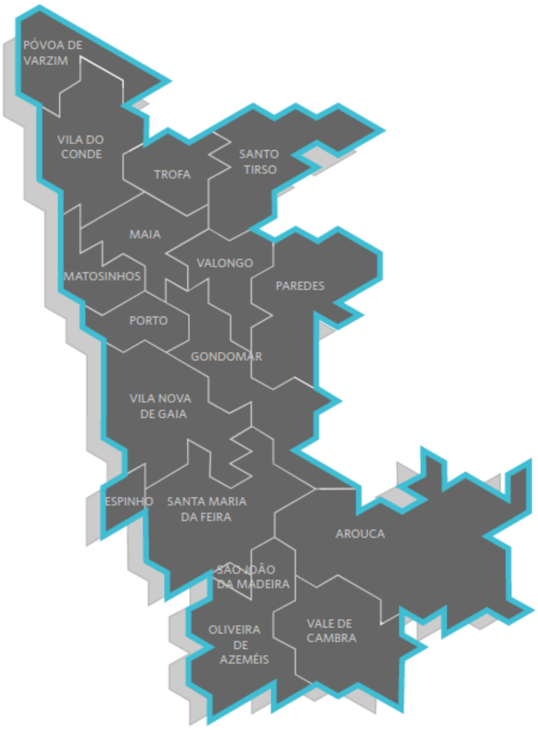


Figure 3.1: Metropolitan Area of Porto

managed by an entity called Transportes Intermodais do Porto (TIP), which is a complementary entity responsible for managing the ticketing system, collecting the fares, and distributing it to the transport operators.

Andante is an entry-only AFC system where a zonal structure defines the fares. The network is divided into geographic travel zones, and the journey fare depends on the number of zones traveled between its origin and destination. There are mainly five types of Andante travel cards:

- **Occasional Ticket** - grants the passenger 1 hour of unlimited uses, for a set of adjacent zones previously chosen when charging the ticket.
- **Monthly Subscription** - grants the passenger an unlimited number of uses for a month in a set of zones chosen by the passenger;
- **Andante 24** - grants the passenger an unlimited number of uses for 24 hours in a set of adjacent zones previously chosen by the passenger;
- **Andante Tour 1** - grants the passenger an unlimited number of uses for 24 hours, in all zones;
- **Andante Tour 3** - grants the passenger an unlimited number of uses for 72 hours, in all zones.

The Andante system works in a time-based manner, allowing passengers to make unlimited journeys in a given time period. Upon starting a journey or changing vehicles, passengers need to tap the travel card on a ticket reader, and the Andante system records each validation.

The passengers can charge their personal smart card at automatic ticket vending machines located in every station. While doing so, they choose the number of adjacent zones they want to cross during a certain journey. The respective fare will increase according to the number of zones included.

3.1.2 AFC Data

This work explores AFC data to understand the mobility profiles of tourists. It relies on data from the AFC system of Porto, Portugal, from the year 2019.

To start a journey in a Metropolitan Area of Porto Public Transport, passengers tap the travel card on a reader and the Andante system records a validation. This system is then responsible for storing data regarding the validation. For each validation, various attributes are being recorded, with those being:

- Travel card ID;
- Validator ID;
- Prodider Code;
- Type of Discount;
- Type of Travel Card;
- Operator;
- Time of Validation;
- Origin Metropolitan Area of Porto Zone;
- Predicted Destination Metropolitan Area of Porto Zone;
- Origin (Station / Bus stop);
- Customer Name, Email and Phone Number for validations on phone.

An approximate total of 82 million validations from the year 2019, was recorded and provided to us so we could achieve our goal.

3.1.3 POI Data

This analysis was complemented with the analysis of a second dataset consisting of the points-of-interest of Porto's city. This was based on Google Places API, which is a service that provides an up-to-date database containing information about various places such as establishments, geographic locations, or points of interest.

To collect this data, we developed a loop that navigates through coordinates in the Metropolitan Area of Porto while applying google maps functions that return the places nearby and their respective attributes.

The service provides several attributes from each place but, for our analysis, the only relevant ones are the **name, category and location coordinates**.

The POIs from the data are classified according to 97 categories, which can be aggregated into 11 main categories [12], with those being:

- **Shopping**, which includes the places categorized as stores, clothing stores, grocery or supermarket, electronics store, jewelry, and shoe store, shopping mall, book and liquor stores;
- **Food**, which includes the places categorized as restaurants, cafes, bakeries and bars;
- **Service**, which includes the places categorized as finance, accounting, car repair, travel agencies, banks, laundries, florists, and insurance agencies;
- **Healthcare**, which includes the places categorized as doctors, dentists, pharmacies and hospitals;
- **Lodging**, which includes the places categorized as lodging;
- **Lifestyle and beauty**, which includes the places categorized as hair care, beauty salons, gyms and spas;
- **Education**, which includes the places categorized as universities and schools;
- **Religion**, which includes the places categorized as churches, cemeteries and places of worship;
- **Leisure and culture**, which includes the places categorized as parks, museums, art galleries, movie theatres, libraries and stadiums;
- **Government**, which includes the places categorized as city hall, courthouses, police and fire stations;
- **Transportation**, which includes the places categorized as transit stations, subway and train stations.

3.2 Methodological Approach

In this subsection, the methodological approach will be described for the data understanding and data preparation CRISP-DM phases as well as the approach used while developing a scoring method for tourist identification and finally, the methodology approach to formalize the spatial, temporal, and spatial-temporal patterns used for pattern characterization.

3.2.1 Data Understanding

A database consisting of Andante's AFC data from the year 2019 was provided for this study, consisting of approximately 82 million validation from 4.4 million different travel cards.

This database consists of 14 CSV files where each file corresponds to a month of the year, with the months' May and October being split into two different files due to their larger size.

Each file stores all validations in the Metropolitan Area of Porto for their respective month, having an average size of 2.5 Gigabytes.

For each validation, various attributes were recorded, which can be seen in table 3.1 showing all columns of the database, and table 3.2, with an example of the columns' attributes for a few validations.

Table 3.1: Database Columns

Name	**Description**	**Type**
Code	Validation ID	Integer
Card	Travel Card ID	Hexadecimal
SNValidator	Validator ID	String
ProviderCode	Provider Code	Integer
Discount	Type of Discount	Integer
TarifCode	Type of Travel Card	Integer
OperatorId	Operator	Integer
Time	Time of Validation	Date-Time
EntryZone	Origin Metropolitan Area of Porto Zone	String
ExitZone	Predicted Destination Metropolitan Area of Porto Zone	String
Station	Origin Station	String
Name	User Name	String
Email	User Email	String
Phone	User Phone Number	Integer

Table 3.2: Example of Validations' Attributes

Code; **Card**; **SNValidator**; **ProviderCode**; **Discount**; **TarifCode**; **OperatorId**; **Time**; **EntryZone**; **ExitZone**; **Station**; **Customer**; **Email**; **Phone**
2219531492; 0x61DFAA748118D23F4D7FF9636784E754; 000081B325; 2; 0; 2; 2; 2019-01-07 17:15:00.000; S8; C1; Mafamude; NULL; NULL; NULL
2227239724; 0x17FF95C158F134324FDF11D8BE548A0E; 00C01DB73B; 3; 29; 2; 5; 2019-01-22 08:05:00.000; S1; S1; NULL; NULL; NULL; NULL
2230365543; 0x49625CD51EE727321B459586E91EDDE5; 000073FBB9; 6; 0; 2; 2; 2019-01-30 00:38:00.000; C1; C10; NULL; NULL; NULL; NULL
2233237615; 0x11A3BF9582BFD827F495AFB87F667CBA; 00C01C2691; 3; 3; 2; 2; 2019-06-29 08:14:00.000; C1; C1; Alto Cruz; NULL; NULL; NULL

3.2.2 Data Preparation

This section presents all the steps performed to clean and transform the raw data before it is analyzed and processed. This section is split into two sub-topics: Data Reduction and Data Cleaning.

3.2.2.1 Data Reduction

During the Data Understanding CRISP-DM phase, it was notable that some of the information in the raw database was redundant and that the data could be reduced to a much smaller and efficient volume.

So, in order to reduce our sample while maintaining similar results, we decided to exclude some columns from the raw data.

Firstly, we decided to cut the columns "Code", "SNValidator", "ProviderCode", "Discount", "OperatorId", "EntryZone" since they were redundant to our study.

After, since each Zone in the Metropolitan Area of Porto occupies such a vast area, we decided not to use the "ExitZone" column.

Finally, we decided to exclude the "Station", "Name", "Email" and "Phone" columns since these attributes were only recorded for validations on the Anda phone app, which had a small sample size.

After this process was completed, we were left with the following columns:

- Card;
- TarifCode;
- Time;
- Station.

3.2.2.2 Data Cleaning

After the data reduction, since there were still inaccurate and unnecessary records, we cleaned the data.

To ensure that each validation was not recorded multiple times, the first step while cleaning our data was to remove all the duplicated validations.

After that, we decided to convert every "Card" attribute from a string to a numeric value to reduce the memory cost.

Finally, in order to make the variables more comprehensible, we decided to rename the columns to more accurate descriptions.

After that, the data preparation phase was completed, and we were left with the Columns seen in Table 3.3.

Table 3.3: Database Columns

Name	Type
CardId	Integer
CardType	Integer
ValidationTime	Date-Time
OriginStation	String

3.2.3 Tourist Identification

For the tourist identification phase, we developed a scoring system that retrieves a card's temporal information and determines if the user is a tourist or not.

To develop the system, we collected the temporal characteristics of tourists by comparing the travel cards most used by tourists (Tour 1 and Tour 3) and the travel cards most used by regular passengers (Monthly Subscriptions).

Every difference will be given a positive or negative weight, and, in the end, all weights are summed up. If the scoring system returns a positive score, the user will be classified as a tourist. Otherwise, he will not be considered as one.

In order to validate this function, the data obtained by estimated tourists was compared with the data from tourists, and the data from non-tourists was compared to the data from regular users.

3.2.4 Pattern Characterization

In the modeling phase, in order to perform a deeper analysis of the information extracted from the mobility profiles of tourists in the city of Porto, we formalized three different patterns, namely, spatial patterns, temporal patterns, and spatial-temporal patterns. We used the K-Means algorithm for the clustering process needed at this stage.

For the **temporal patterns**, the data was characterized into temporal travel patterns based on the day, time of day, and season of the year of the validations. This allowed us to identify for non-frequent passengers the busiest hours, days of the week, seasons of the year, and much more.

With the **spatial-temporal patterns**, it was possible to profile the stations' public transport daily usage. These patterns allowed us to identify how much tourists congest a station for a specific time.

Finally, for the **spatial patterns**, the data was clustered into spatial travel patterns based on the location and the points of interest within walking distance of that location. With these patterns, it was possible to identify the most frequented locations by tourists and to characterize each station.

So we could verify the validity of each generated cluster, they underwent a validation process. While performing the K-Means clustering, we used the elbow method with distortion in order to select the optimal number of clusters. This allowed us to select the clusters with the smallest intra-cluster distance (distance between two objects of the same group) and the largest inter-cluster distance (distance between objects that belong to different groups), in order to validate the choice of the number of clusters.

Chapter 4

Tourists Detection

This section demonstrates the method we developed to identify users as tourists through their travel cards' validations.

We initially had the idea of using only the data from Tour1 and Tour3 ticket cards but, by doing that, we were excluding a large amount of data from occasional tickets, which are also used by tourists. To solve this, we decided to implement a scoring system that identifies a user as a tourist or not considering all the validations made with their travel card.

While developing the scoring system, we used the Tour1 and Tour3 cards' data as a representation of tourists since they are designed for these users and the Monthly Subscriptions' data to represent regular users since its features attract local users. Statistics regarding both types of cards are shown in Table 4.1.

Finally, the method developed was applied to the remaining ticket data to identify potential tourists.

Table 4.1: Tour and Monthly Subscription Cards' Statistics

	Nº Validations	**Nº Cards**
Tour Cards	1120165	153824
Monthly Subscriptions Cards	140239493	374090

4.1 Pre-Processing

The first step while preparing the data was to generate two databases, one for the Tour cards and another for the Monthly Subscription Cards, where we only needed to use the columns with information related to the Card and Time of Validation.

Each validation was grouped by its card id, and for each card, several attributes were calculated based on its temporal information and sequences that could help to identify a tourist. For this study, we defined sequences as consecutive days using the travel cards. As soon as the card goes one day without registering a validation, the sequence ends.

The attributes calculated for each card were the following:

- **TripsDayWeek** - Array<Int>
  - Stores the number of validations recorded for each day of the week where Index 0 of the Array corresponds to Monday.
- **TripsHourWeekday** - Array<Int>
  - Stores the number of validations recorded in weekdays (Monday-Friday) for each hour of the day where the time-frame [00:00,01:00[corresponds to Index 0 of the Array and so on until the [23:00,24,00[time-frame.
- **TripsHourWeekend** - Array<Int>
  - Stores the number of validations recorded in weekends (Saturday,Sunday) for each hour of the day where the time-frame [00:00,01:00[corresponds to Index 0 of the Array and so on until the [23:00,24,00[time-frame.
- **TripsMonth** - Array<Int>
  - Stores the number of validations recorded for each month of the year where Index 0 of the Array corresponds to January.
- **NumberSequences** - Int
  - Number of sequences identified in the card.
- **NumberDaysPerSequence** - Array<Int>
  - Stores the number of consecutive days the travel card was used for each sequence.
- **NumberTripsPerSequence** - Array<Int>
  - Stores the number of validations recorded for each sequence found.
- **HasCycle** - Boolean
  - Boolean that is True if the interval between each sequence is systematic or, in other words, if the interval between sequences is consistent.
  - Number of sequences >= 2.
  - Interval (standard deviation of 1) must appear at least 50% of the time.
  - Pseudo-Code : $Count(Interval\ or\ Interval \pm 1) \div Count(All) \geq 0.5$

4.2 Results obtained for each variable

In this subsection, we compared the results for both types of cards by summing all validation variables for each type of user.

Since the amount of validations from Monthly Subscriptions is significantly higher than the Tour cards' validations, we also normalized these results, making it a fairer comparison.

To perform such comparison, the results were converted into percentages relative to the probability of occurrence for the respective card type. Every percentage resulting from tourist cards is equal to 100%, and the same applies to the results from Monthly Subscription cards.

4.2.1 TripsDayWeek

Since inhabitants of a city and tourists follow quite different weekly schedules, we decided to calculate the number of trips for each day of the week as it could help to determine a tourist through their validations.

Through an analysis of the bar chart in Figure 4.1, it is possible to discern that the number of trips is relatively constant for tourists throughout the whole week. In contrast, it is steady for regular users from Monday to Friday but has a dip on Saturday and Sunday.

While a weekend trip amongst tourists is much more common than amongst regular users, in contrast, the opposite is evident for weekdays.

In addition, a slight growth of validations by tourists is noted throughout the week, reaching its maximum on Friday. However, for both types of passengers, Sunday is the day with the lowest percentage of validations.

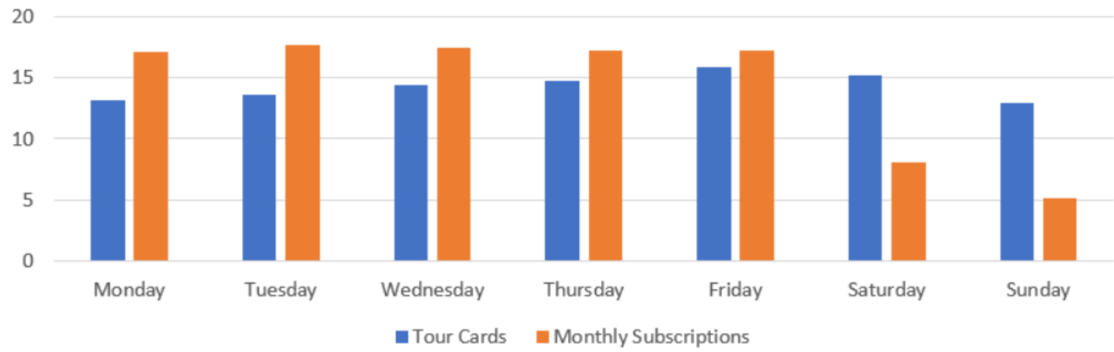


Figure 4.1: Percentage of validations for each day of the week based on Table A.1

4.2.2 TripsHourWeekday

Considering citizens and tourists' considerably different daily schedules, we also decided to calculate the number of trips for each hour of the day to assist in discovering a tourist through their validations.

For this study, we initially stored all validations for each hour of the day but ended up dividing this into weekdays and weekends due to most users having different schedules for these periods.

After adding up all variables for each user type, we proceeded to compare the results from Tourists and Regular Users for weekdays and, with that, we realized that the number of validations from Regular users registered a peak during rush hours whereas the validations from Tourists are more spread across the day and even lower in this rush hours.

This can be seen in the line chart in Figure 4.2.

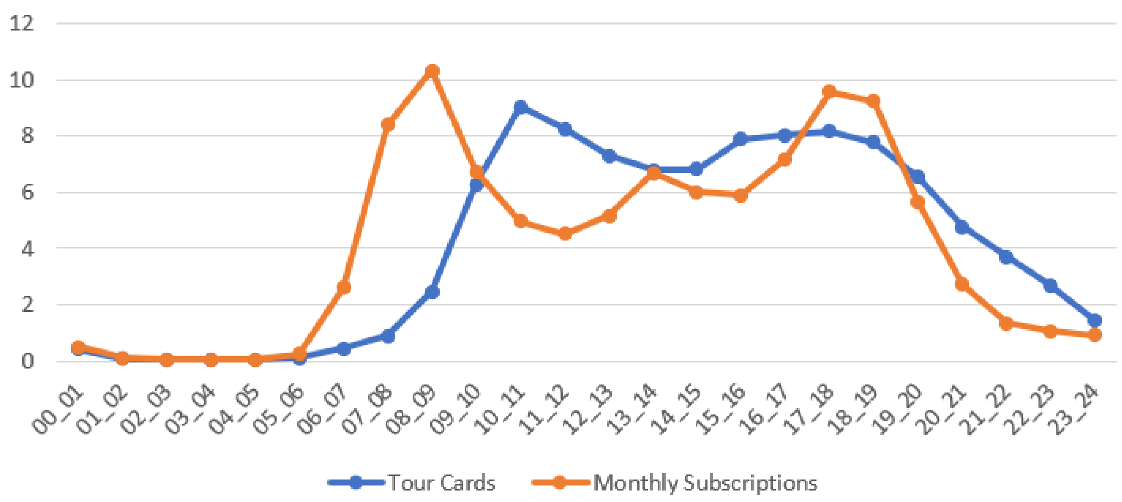


Figure 4.2: Percentage of validations, from Monday to Friday, for each hour of the day based on Table A.2

4.2.3 TripsHourWeekend

From the Chart in Figure 4.3, we can see that for the TripsHourWeekend the difference in usage percentage of each hour from Tourists and Regular Users is much more insignificant than for the TripsHourWeekday variable. Or, in other words, the daily movement of tourists and Regular Users is much more similar during weekends than on weekdays.

While the normalized totals from Tourists on weekends and weekdays are very similar throughout the whole day, this is not the case for the Regular Users. On weekdays there is much more emphasis on rush hours (07:00 to 09:00 and 17:00 to 19:00) than on the weekends.

This was expected since working days in Portugal range from Monday to Friday, and it verified our decision to divide the daily activity into weekdays and weekends.

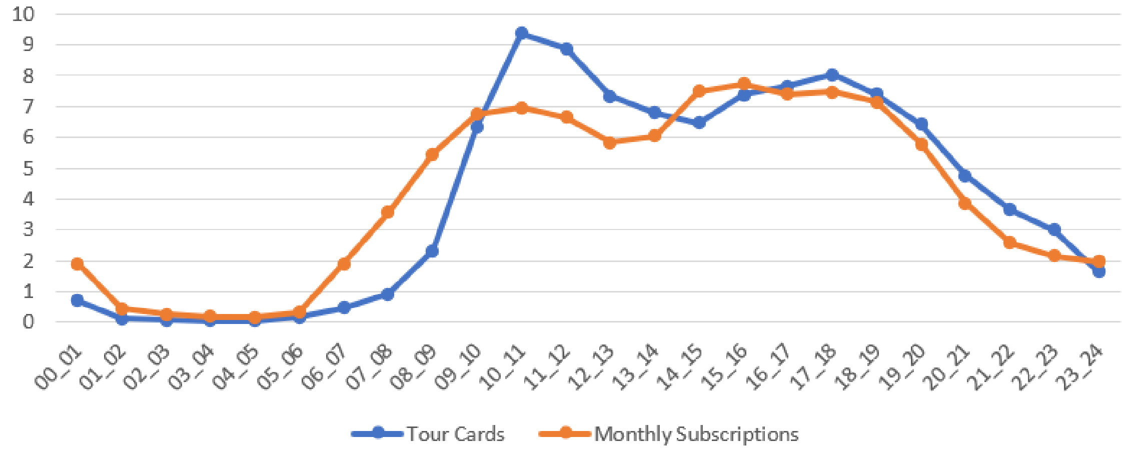


Figure 4.3: Percentage of validations, from Saturday and Sunday, for each hour of the day based on Table A.3

4.2.4 TripsMonth

To help the identification of the seasons and months with the highest number of tourists, we counted the number of validations per month.

From the TripsMonth variable, we can see from the graph in Figure 4.4 that the percentage of Tourists validating their ticket is higher from April until October, which corresponds to the summer and the most tropical weather periods in Porto. In contrast, Regular Users register validations more consistently throughout the year, with a slight increase from October to December and in May.

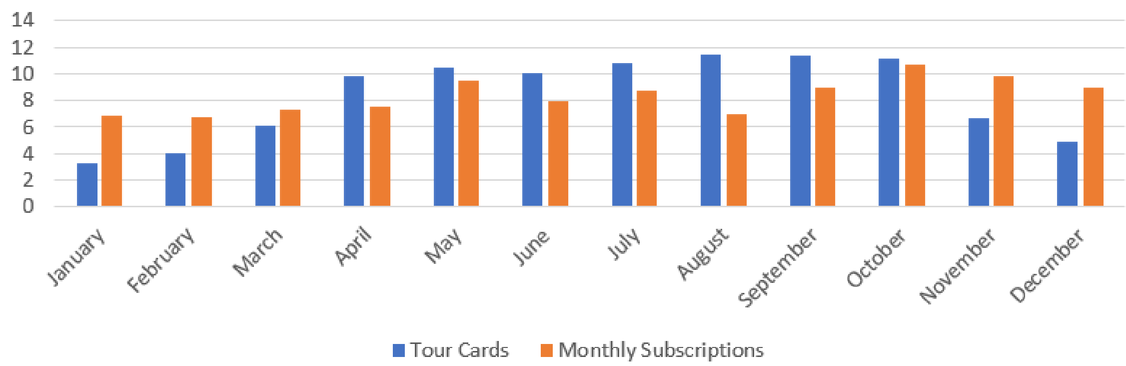


Figure 4.4: Percentage of validations for each month of the year based on Table A.4

4.2.5 NumberSequences

In our work, we assume that a tourist would use public transports in fewer sequences than a regular user, so we used a variable to count the number of sequences registered in a card.

For the NumberSequences variable, it can be seen from the Line Chart in Figure 4.5 that the percentages remained somewhat consistent (around 7-9%) for Regular Users. In contrast, most Tourists (92.47%) registered a single sequence, with the number of users that formed more than two sequences being insignificant due to their low value.

Due to this discrepancy of values from Tourists and Regular Users the NumberSequences variable carried a lot of weight for our scoring system to determine if a public transport traveller is a Tourist from their validations.

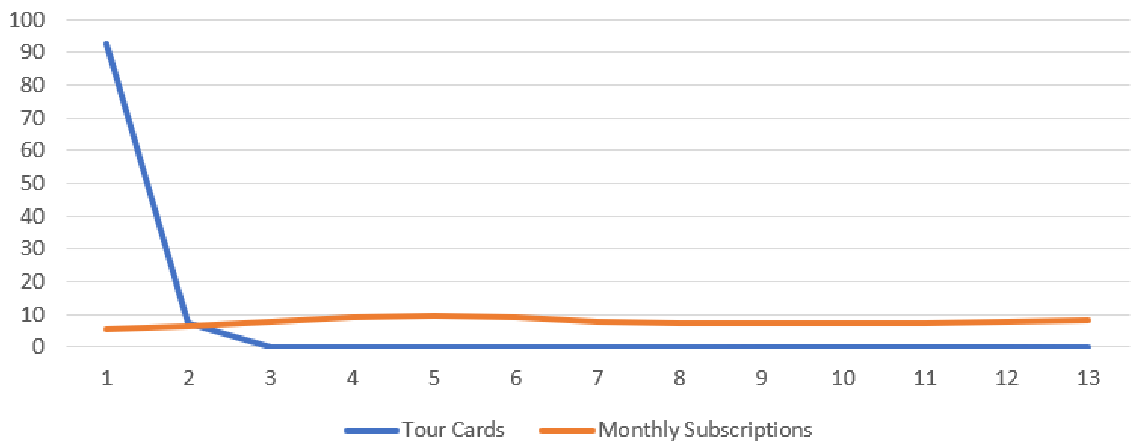


Figure 4.5: Percentage of Number of Sequences for each card based on Table A.5

4.2.6 NumberDaysPerSequence

We assume in this work that a tourist does not stay long in the city he is visiting. Therefore, the variable NumberDaysPerSequence registers the number of days in a sequence to aid in the detection of tourists.

From the total results obtained, we noticed that the lower numbers of consecutive (1-4) days recording validations favor tourists while the opposite occurs for longer sequences (5+ days). This can be seen from the Line Chart in Figure 4.6

But, due to the characteristics of the Tour1 cards(24h) and the Tour3 cards(72h), these results were expected. However, to further substantiate these results, Gössling et al. [16] revealed that the average length of stay in Portugal from the year 2015 is from 3 to 4 days which coincides with 2 of the most common values obtained for Tourists' sequences.

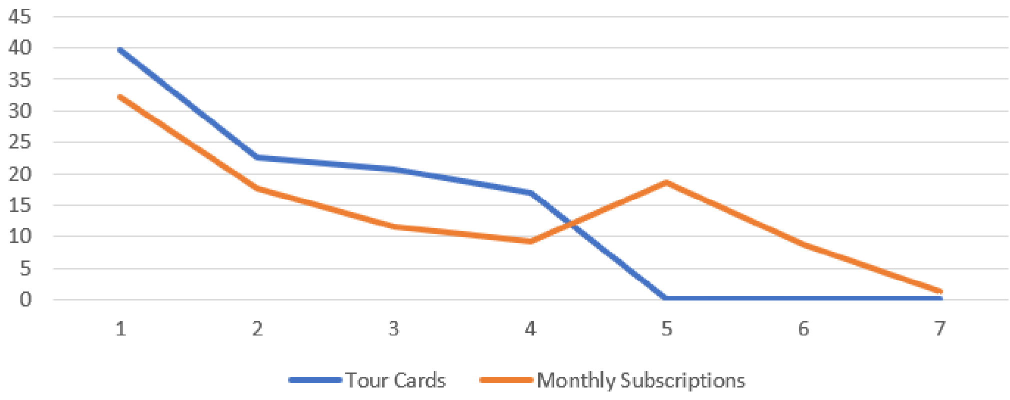


Figure 4.6: Percentage of Number of consecutive days per sequence based on Table A.6

4.2.7 NumberTripsPerSequence

Another variable we decided to use that could help determine the type of user using a travel card was the NumberTripsPerSequence. But, to our surprise, after getting all sequences and their respective number of trips, we noticed that the results were a bit inconsistent between values and quite similar for both types of users. This is notable on the line graph in Figure 4.7, where it shows that both lines have similar curvature despite the Monthly Subscriptions fluctuating more between values.

Regarding the number of trips per sequence, it was expected that it would decrease exponentially, as was the case for both types. However, it was noticeable that the percentages for even numbers of trips were higher than for the odd ones, but only for regular users. This may have been the case since, usually, a tourist travels around the city from place to place until he returns to his origin, while a regular user often makes routine trips from home to a destination and back, which corresponds to an even number of trips.

Due to all that, we decided not to use this variable for our scoring system.

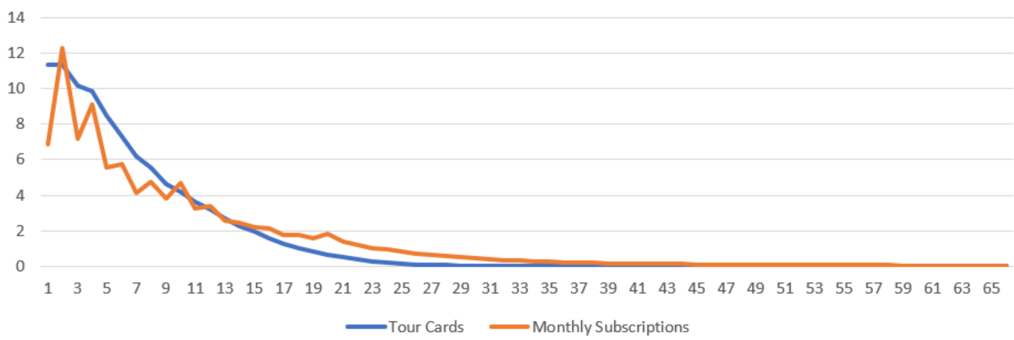


Figure 4.7: Percentage of Number of trips per sequence based on Table A.7

4.2.8 HasCycle

When developing our scoring system, we found it useful to introduce a variable that registers if the sequences recorded follow a pattern, which may be more common for people living in the city. (ex: Validation every Monday). Therefore, the HasCycle variable was introduced.

After getting the total results, we soon discovered that this was the variable that showed the most considerable distinction between the types of users in the study (Figure 4.8). Whereas 37.25% of Monthly Subscriptions cards contained a cycle, only 3 out of 153821 (0.002%) Tour cards contained a cycle in them. So, if a card has a cycle, it is highly probable to be a Regular User.

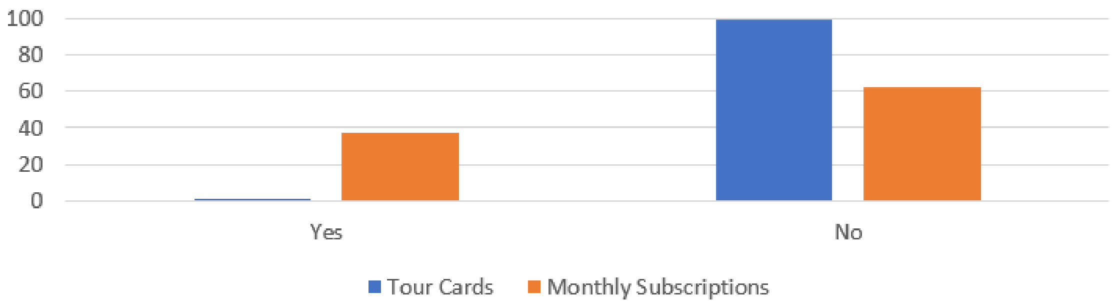


Figure 4.8: Percentage of cards with cycles based on Table A.8

4.3 Scoring System

This section describes the system/function responsible for identifying tourists and non-tourists from travel card validations.

4.3.1 Arguments

After analyzing every variable mentioned, we opted to use the following variables in our scoring system for each card:

- TripsDayWeek;
- TripsHourWeekday;
- TripsHourWeekend;
- TripsMonth;
- NumberSequences;
- NumberDaysPerSequence;
- HasCycle.

4.3.2 Attribute Weights

The next step was to attribute the correct weight for each value in each variable since they showed different results from the study.

Firstly, since the number of results for some values is near to zero for tourist users, we assumed that if a card presents any of these following values, the user is automatically ruled as a non-tourist.

- **HasCycle = True**, since only 3 out of 153824 tour cards contained cycles.
- **NumberSequences > 1**, since only 25 out of 153824 tour cards had more than 1 sequence.
- **NumberDaysPerSequence > 7**, since 0 sequences from tour cards was higher than 7 straight days.

As for the other values in each variable, in order to attribute their respective weight, we followed the Equation 4.1.

$$Score = Integer(Tourists'Percentage(Value) - RegularUsers'Percentage(Value)) \tag{4.1}$$

As an example, Equation 4.2 shows the score for the variables TripsDayWeek, TripsHourWeekday and TripsMonth for a card that registered a validation on Monday at 8:30 in August, where TP and RP correspond to "Tourists' Percentage" and "Regular Users' Percentage", respectively.

$$\begin{aligned}
Score(TripsDayWeek) &= Integer(TP(Monday) - RP(Monday)) \\
&= Integer(13.19 - 17.16) \\
&= -4 \\
Score(TripsHourWeekday) &= Integer(TP(08:00-09:00) - RP(08:00-09:00)) \\
&= Integer(2.46 - 10.35) \\
&= -8 \\
Score(TripsMonth) &= Integer(TP(August) - RP(August)) \\
&= Integer(11.46 - 6.95) \\
&= 5
\end{aligned} \tag{4.2}$$

By applying Equation 4.1 to all values of each variable we obtained the weights and conditions in Equation 4.3.

$$\begin{aligned}
TripsDayWeek &= [\text{-}4,\text{-}4,\text{-}3,\text{-}2,\text{-}1,7,8] \\
TripsHourWeekday &= [0,0,0,0,0,0,\text{-}2,\text{-}8,\text{-}8,0,4,4,2,0,1,2,1,\text{-}1,\text{-}1,1,2,2,2,1] \\
TripsHourWeekend &= [\text{-}1,0,0,0,0,0,\text{-}1,\text{-}3,\text{-}3,0,2,2,2,1,\text{-}1,0,0,1,0,1,1,1,1,0] \\
TripsMonth &= [\text{-}4,\text{-}3,\text{-}1,2,1,2,2,5,2,0,\text{-}3,\text{-}4] \\
NumberSequences &= [87,0] \\
NumberDaysPerSequence &= [8,5,9,8,\text{-}19,\text{-}9,\text{-}1] \\
if(HasCycle) &\rightarrow NotTourist \\
if(NumberSequences > 2) &\rightarrow NotTourist \\
if(NumberDaysPerSequence > 7) &\rightarrow NotTourist
\end{aligned} \tag{4.3}$$

Regarding the conditions and weights calculated, these instances are directly related to the data generated by the validations registered in the Metropolitan Area of Porto. Due to the different movement patterns in different locations, using this algorithm with different data may not reveal such accuracy. Therefore, it would be conventional to recalculate the variables percentages for new AFC data.

4.3.3 Function Features

We decided to set a maximum and minimum cap on their scores for the variables with an unrestrained total score (due to being arrays without a fixed size). Setting a cap of -30 to 30 was enough to make sure that these variables were influencing the final score without overwhelming the other variables.

Regarding the function, it follows the following format:

1. Receives, as input, all variables;
2. Tests if at least one of the conditions is occurring and, if so, the function will return False, i.e., the user is not a tourist;
3. Calculates the score for each variable:
   - If the variable is an array, the score is calculated by doing a wise-product with its respective weight array. The score is then normalized not to exceed the maximum and minimum cap.
   - If the variable is an integer, the score corresponds to the weight's array element, which index is that integer.
4. Sums all the scores calculated and, for a user to be considered a tourist, that total must be greater than 0. Otherwise, the user is considered a non-tourist.

4.3.4 Pseudo-Code

The following Algorithm 1 represents the pseudo-code of the function that predicts the user type, returning True if he is a tourist or False if not.

Algorithm 1 Calculates if a card was used by a tourist or not.

Input: TripsDayWeek, TripsHourWeekday, TripsHourWeekend, TripsMonth, NumberSequences,NumberDaysPerSequence, HasCycle

```
score ← 0
scoreTripsDayWeek ← [-4,-4,-3,-2,-1,7,8]
scoreTripsHourWeekday ← [0,0,0,0,0,0,-2,-8,-8,0,4,4,2,0,1,2,1,-1,-1,1,2,2,2,1]
scoreTripsHourWeekend ← [-1,0,0,0,0,0,-1,-3,-3,0,2,2,2,1,-1,0,0,1,0,1,1,1,1,0]
scoreTripsMonth ← [-4,-3,-1,2,1,2,2,5,2,0,-3,-4]
scoreNumberSequences ← [87,0]
scoreNumberDaysPerSequence ← [8,5,9,8,-19,-9,-1]
if hasCycle or NumberSequences > 2 or max(NumberDaysPerSequence) > 7 then
   Return False
end if
score ← score + MinMax(SumArray(MultiplyIndex(scoreTripsDayWeek, TripsDayWeek)))
score ← score + MinMax(SumArray(MultiplyIndex(scoreTripsHourWeekday, TripsHourWeekday)))
score ← score + MinMax(SumArray(MultiplyIndex(scoreTripsHourWeekend, TripsHourWeekend)))
score ← score + MinMax(SumArray(MultiplyIndex(scoreTripsMonth, TripsMonth)))
score ← score + scoreNumberSequences[NumberSequences]
score ← score + MinMax(SumArray(scoreNumberDaysPerSequence, NumberDaysPerSequence)))
if score > 0 then
   Return True
end if
Return False
```

The functions invoked in the Pseudo-Code were:

- **SumArray**, sum of all numbers inside an array.
- **MultiplyIndex**, element-wise product of two arrays.
- **MinMax**, sets a 30 maximum and -30 minimum limit to an integer, converting the integer to the limit it/if exceeds.

4.4 Discussion Results

In this section, we present the results obtained after the scoring system identifies each card as a tourist or not. From the confusion matrix, we are able to observe the error rate and, consequently, the function's accuracy. In the Card Type Distribution subsection, we illustrate how the function performed on the different card types.

4.4.1 Confusion Matrix

In order to test the accuracy of the scoring system, since we decided to use the Tour1 and Tour3 cards' data as a representation of tourists, and the Monthly Subscriptions' data to represent regular users, we calculated a Confusion Matrix where:

- **True Positives:** Tour Cards, Predicted Tourist
- **True Negatives:** Regular User, Predicted Not Tourist
- **False Positives:** Regular User, Predicted Tourist
- **False Negatives:** Tour Cards, Predicted Not Tourist

As the Confusion Matrix in table 4.2 shows, we obtained a Total Accuracy of 98.52% and an Error Rate of 1.48% while comparing cards from Tourists and Regular Users. Although this does not directly correspond to the total accuracy as we are not using data from the city's occasional users, it was a helpful model to follow while developing the function and assigning weights.

Table 4.2: Confusion Matrix for Scoring System

		Predicted Values		
		Torurist	Not Tourist	Accuracy
Real Values	Tour Cards	152790	1035	99.33%
	Monthly Subscriptions	6787	367304	98.19%
	Accuracy	95.75%	99.72%	**98.52%**

		Predicted Values	
		Positive	Negative
Real Values	True	28.94%	0.20%
	False	1.28%	69.58%

4.4.2 Card Type Distribution

Finally, in order to get a more detailed view of our scoring method, we decided to group the results in four different card types, with those being **Monthly Subscriptions**, **Occasional Tickets**, **Tour Cards** (Tour1 and Tour3) and **24H Cards** (Andante 24).

From the statistics in Table 4.3, we drew the following conclusions:

- The number and percentage of tourists for the Tour Cards (99.33%) and Monthly Subscriptions (1.81%) seem plausible, which was expected since these were the cards used as our test subjects.
- The results from occasional tickets appear to be higher than expected (63.48%). This happened because we based a citizen's behavior on Regular Users, not including citizens with occasional or irregular behavior. Since the behavior of tourists and occasional users may be similar, the score function will not have the necessary information to separate these types of users accurately.

- As for the 24h riders, since this card has very similar characteristics to the Tour1 Cards (equal limitations, varying only in zoning), it was expected for the number, and consequently, the percentage of tourists is as high as it is (63.67%).

- Finally, due to the vast amount of cards being occasional tickets, the total percentage of cards used by tourists in 2019 predicted was higher than expected (59.53%). Although the model developed worked very well while distinguishing tourists from regular users, since there were challenges when identifying tourists from occasional users, this higher percentage transpired.

Table 4.3: Number of predicted card types by Scoring System

	Tourist	**Total**	**Tourist Percentage**
Monthly Subscriptions	6787	374091	1.81%
Occasional Ticket	2391515	3767343	63.48%
Tour Cards	152790	153825	99.33%
24H Cards	91462	143660	63.67%
Total	2642554	4438919	59.53%

Chapter 5

Pattern Formalization

This section presents some characteristics and patterns for cards utilized by users classified as tourists from our scoring system.

We described Temporal characteristics through Daily, Weekly, and Monthly Patterns. Then, we clustered the stations located in the Metropolitan Area of Porto with their temporal attributes to form a Spatio-Temporal Pattern. Finally, by crossing the data with another database containing the points of interest in the city of Porto, we clustered the stations based on their location, obtaining Spatial Patterns.

5.1 Temporal Characterization

After getting the tourists by the scoring system through their validations, we proceeded with a temporal analysis concerning these users.

For the temporal characterization, we decided to portray the daily, weekly, and monthly patterns since they could be used to represent the busiest periods for tourism in the city of Porto.

Like the scoring system, we divided the "ValidationTime" variable into several temporal-related variables stored for each card. With these variables, we were able to obtain the desired patterns.

5.1.1 Weekly Patterns

To form the weekly patterns, we decided to store the number of validations on each day of the week for each card (same process as the "TripsDayWeek" variable mentioned in the scoring system). Table 5.1 shows the format used while storing this information in each card.

Table 5.1: Approach used while storing data for Weekly Patterns

CardId	Mon	Tue	Wed	Thu	Fri	Sat	Sun
id1	2	0	0	0	0	3	7
id2	0	1	3	2	1	0	0
...							

5.1.2 Daily Patterns

For the daily patterns, we stored, for each card, 48 columns with the number of trips per hour of weekdays and weekends, which was a concatenation of the "TripsHourWeekday" and the "TripsHourWeekend" variables of our scoring system. Table 5.2 displays the setup adopted while filling in this data for each card with H_0 representing weekdays and H_1 the weekend.

Table 5.2: Approach used while storing data for Daily Patterns

CardId	$H_0 0$	$H_0 1$	...	$H_0 23$	$H_1 0$	$H_1 1$	...	$H_1 23$
id1	0	0	...	0	2	1	...	0
id2	0	1	...	2	0	0	...	0
...								

5.1.3 Monthly Patterns

As for the monthly patterns, it was registered the cards' number of validations for each month. (same process as the "TripsMonth" variable mentioned in the scoring system). Table 5.3 shows the format used while storing this data per card.

Table 5.3: Approach used while storing data for Monthly Patterns

CardId	Jan	Feb	Mar	...	Oct	Nov	Dec
id1	0	0	0	...	0	10	0
id2	0	1	6	...	0	0	0
...							

5.1.4 Discussion Results

In this section, we will present all the temporal patterns formed from tourist cards and contextualize them by comparing to the results obtained from the Tour cards and Monthly Subscriptions mentioned in Chapter 5.

5.1.4.1 Weekly Pattern

Figure 5.1 shows the number of validations for each day of the week from users estimated as Tourists and Non-Tourists. To contextualize these results, we can compare them to the results from Tour Cards and Monthly Subscriptions displayed in Figure 4.1.

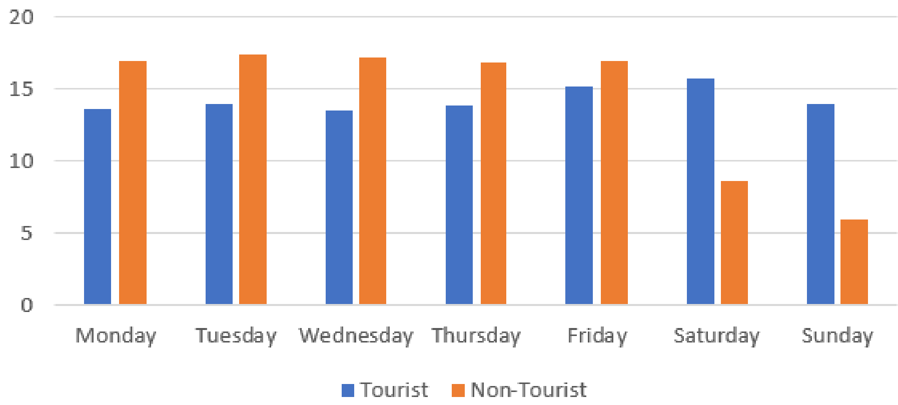


Figure 5.1: Percentage of validations for each day of the week based on Table B.1

From this comparison, we observed that:

- The results from Tourist Cards are similar to those from Tour1 and Tour3 cards;
- The results from Non-Tourist cards are identical to the Monthly Subscriptions.

Since our case study for tourists (Tour Cards) only consists of a small percentage of predicted tourists (5.78%) and our case study for the non-tourists is only 20.45%, we can conclude that the variable used to measure the number of trips ("TripsDayWeek") is having the desired effect.

Finally, we can also recognize that individual card users (occasional non-tourist users) and tourists share similar weekly patterns, which may explain the system's difficulties in separating these users while also validating the system's accuracy.

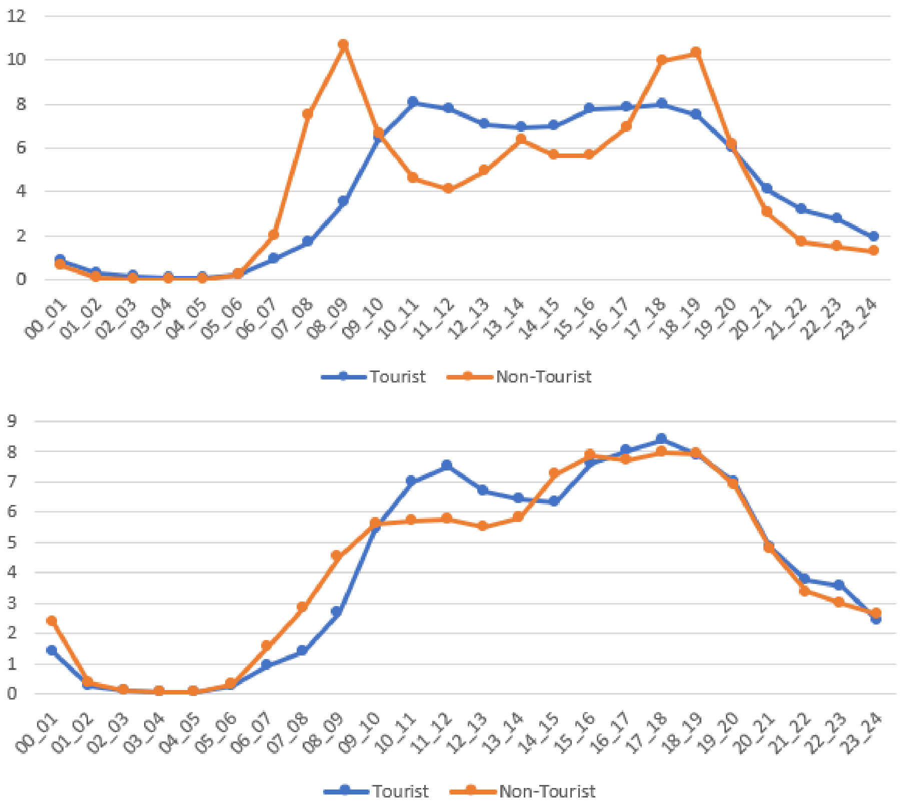
5.1.4.2 Daily Pattern

As we did for the weekly patterns, we decided to compare the percentage of trips per hour of the day (weekend and weekdays) from tourists and non-tourist (shown in Figure 5.2) with the results from the Tour cards and Monthly subscriptions in Figures 4.2 and 4.3.

From these results, we can see that:

- Results from tourists and Tour cards are alike for both weekdays and weekends. The only minimal discrepancy is during and after the morning rush hours (07:00 to 10:00) for both weekdays and weekends, where there is a higher percentage of validations from predicted tourists in rush hours and lesser after that period (variation of 1% in each hour).
- Non-tourists' and Monthly Subscriptions' results are identical.

Since the results compared were moderately alike, we can perceive that the "TripsHourWeekday" and "TripsHourWeekend" variables conditioned the final result of the scoring system, which helps in validating the accuracy of the scoring method.



(a)

(b)

Figure 5.2: Percentage of validations for each day of the week based on Table B.2. **(a)** Weekdays. **(b)** Weekend.

5.1.4.3 Monthly Pattern

To finish the temporal analysis, we also compare the results obtained, for each month, from tourists and non-tourists with those obtained from tourist cards and monthly subscriptions. Figure 5.3 shows the first two types mentioned, while Figure 4.4 shows the others.

With these results, we were able to observe that:

- The monthly results of users predicted as non-tourist are very similar to those from monthly subscriptions, with both being relatively constant throughout the year. The only disparity occurred in April, where the percentage of validations from predicted Non-Tourists was significantly higher than the percentage from Monthly Subscriptions (9.85% and 7,52%).

- Although there is still a higher incidence during the most tropical seasons for both predicted tourists and Tour travel cards, this incidence is more prominent for the Tour cards. The results are varied, with an average difference of 1.22% for each month of the year.

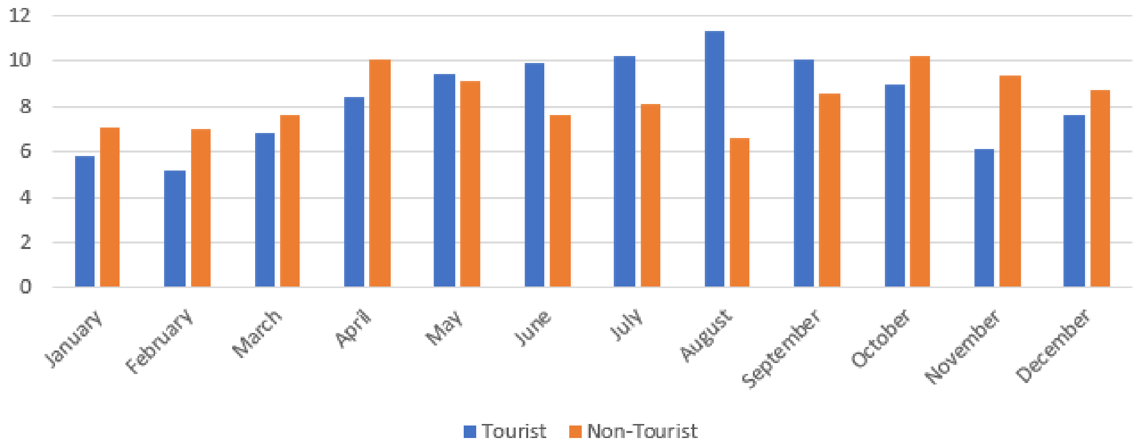


Figure 5.3: Percentage of validations for each month of the year based on Table B.3

Given the results obtained, we prognosticate that the distribution of monthly validations could be a crucial characteristic while distinguishing occasional users that live in the city from tourists. This happens to be the case since there are periods throughout the year that are more favorable for tourism, while occasional users display a monthly distribution more distributed throughout the year.

5.2 Spatial-Temporal Characterization

For the Spatial-Temporal characterization, we decided to cluster the stations based on their validations' daily distribution. Since the database provided to us did not include the Origin Station for Bus validations (only the zone, which occupies a large area), we decided to only perform a characterization for Train and Railway stations.

In order to obtain each cluster of stations, we performed the following data mining techniques: Pre-process the AFC data; Calculate the optimal number of clusters; Apply the K-Means clustering algorithm. These processes will be detailed in more depth in the following sections.

5.2.1 Pre-Processing

Our first step while preparing the data was to select only the columns necessary for our spatial-temporal characterization, which were the following columns:

- "OriginStation" <Sting>;
- "ValidationTime" <Date-Time>.

Then, we grouped the AFC data by origin station to obtain, for each station, the temporal information of each validation performed at the station. With this information, we counted the number of trips for each hour of the day (weekday and weekend), finally obtaining a database with the format presented in table 5.4.

Table 5.4: First approach used while storing data for the Spatial-Temporal characterization

Origin Station	H_00	H_01	...	H_023	H_10	H_11	...	H_123
Station1	1050	650	...	10010	200	40	...	1500
Station2	20	6	...	130	7	5	...	100
...								

Finally, so that the total number of validations is not the only constraint for the clustering process, we normalized the columns that hold the number of trips for each hour. To do such, we transformed each of the variables into the range [0,1] by following Equation 5.1.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{5.1}$$

After the normalization, the first format shown in Table 5.4 was directly converted to the format in Table 5.5.

Table 5.5: Final approach used while storing data for the Spatial-Temporal characterization

Origin Station	H_00	H_01	...	H_023	H_10	H_11	...	H_123
Station1	0.1044	0.0645	...	1.0	0.0195	0.0035	...	0.1494
Station1	0.0015	0.0001	...	0.0125	0.0002	0.0	...	0.0095
...								

5.2.2 K-Means

With the database now prepared, our next step was to generate clusters of stations to group stations with similar daily distributions. To do so, we decided to use the K-Means algorithm.

The K-Means is a popular clustering algorithm due to its simplicity and suitability for quantitative variables. It is an iterative process that divides observations into k clusters by calculating the Euclidean distance between the observation and the centroid(cluster center) and assigning each observation to its nearest centroid.

The algorithm will be applied to all 48 variables that store the number of validations for each hour of the day (weekday and weekend) throughout the year. However, since the number of clusters is a parameter to be provided by the user, we needed to find the optimal number of clusters before the K-Means algorithm was applied to our data.

5.2.3 Number of Clusters

To determine the optimal number of clusters for our data, we opted to use the **Elbow Method**. This method consists of plotting the values produced by different numbers of clusters and picking the point that resembles an elbow on the curve as the number of clusters to utilize.

As our scoring metric, we used the distortion, which measures the aggregate of squared distances from each point to its assigned centroid. From that, we can recognize that a lower distortion and consequently higher number of centroids is favorable, as the observations in each cluster continue to be more similar to one another the bigger the K.

From the elbow curve obtained (displayed in Figure 5.4), we decided to select six as our number of clusters since the difference in distortion score between six and seven clusters was only 0.4.

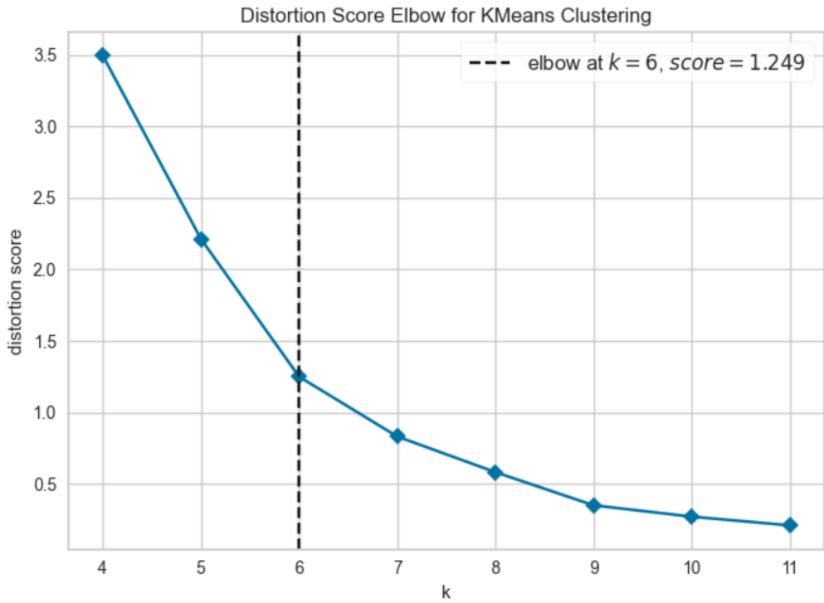


Figure 5.4: Elbow method result with distortion scores

5.2.4 Discussion Results

In this section, we will present all the spatial-temporal patterns formed by clustering the stations based on their validations' daily distribution.

5.2.4.1 Clusters Formed

By executing the K-Means algorithm with 6 clusters on our database, the following clusters was generated:

- **Cluster 1**: Aeroporto;
- **Cluster 2**: Trindade;
- **Cluster 3**: São Bento, Campanhã, Casa da Música, Bolhão, 24 de Agosto;

- **Cluster 4**: João de Deus, Estádio do Dragão, H. São João, Aliados, Santo Ovídio, Srª da Hora, Marquês, Sete Bicas, Jardim do Morro;

- **Cluster 5**: 25 stations that registered a relatively low number of validations per hour, in comparison with the clusters 1 though 4;

- **Cluster 6**: 96 stations that registered almost no validations.

From these results, we can see that even though we normalized the number of validations, the "Aeroporto" and "Trindade" stations had such a higher amount of validations for each hour that they were separated in their own cluster.

For the first cluster, the fact that the station near the Airport has such an immense amount of validations per hour may help in proving the efficiency of our method on separating tourists from non-tourists.

As for the second cluster, since the station in question is the central Transit hub station in the Metropolitan Area of Porto, it would be expected to be in a cluster of its own.

5.2.4.2 Daily Distribution

Firstly, since cluster 5 and cluster 6 had a lower amount of validations, we decided to focus our study on the first 4 clusters.

To demonstrate the daily distribution for each cluster during weekdays and weekends, we generated line charts that plot the number of validations for each hour of the day in each station of a cluster. This can be seen in Figure 5.5.

Such information can be applied to increase the tourists' appeal and explore business opportunities for these specific hours and areas. There are many examples of this, but some of them could be:

- Concentrating more taxis and Uber cars at the busiest hours per station;

- Parking food trucks near stations where there is a higher amount of validations during regular eating schedules (lunch and dinner);

- Decreasing the waiting time for public transportation during rush hours by increasing the number of vehicles en route throughout that period;

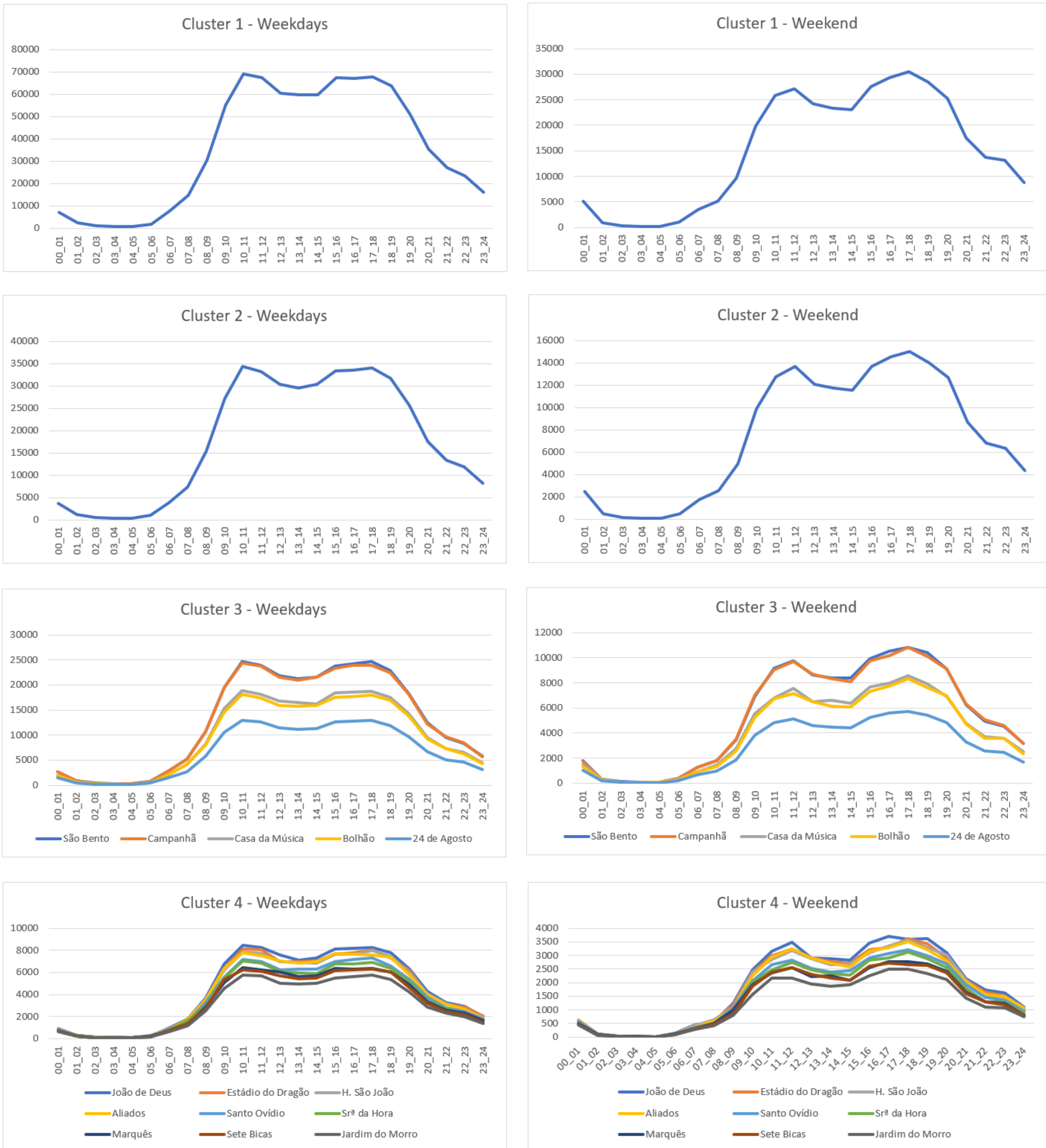


Figure 5.5: Daily distribution for clusters based on Tables B.4 and B.5

5.3 Spatial Characterization

Finally, for the Spatial characterization, we clustered the stations from the first four clusters in the spatial-temporal characterization based on the points of interest near them.

The processes performed and deeply detailed in this section were the following: Collect all the POI in the Metropolitan Area of Porto; Pre-process the AFC data; Calculate the optimal number of clusters; Apply the K-Means clustering algorithm; Discussion of Results.

5.3.1 POI Data

The first step while preparing the spatial characterization data was to collect all the points of interest located in the city of Porto. To do such, we used the Google Places API for the whole Metropolitan Area of Porto.

To do such, we developed a loop that goes through latitudes and longitudes inside the Metropolitan Area of Porto and, inside that loop, we used the "places_nearby" function that returns a list of the 20 nearest places to a coordinate and the "places" function to obtain the place's attributes. The Latitude and Longitude we used to represent the area in question was the following:

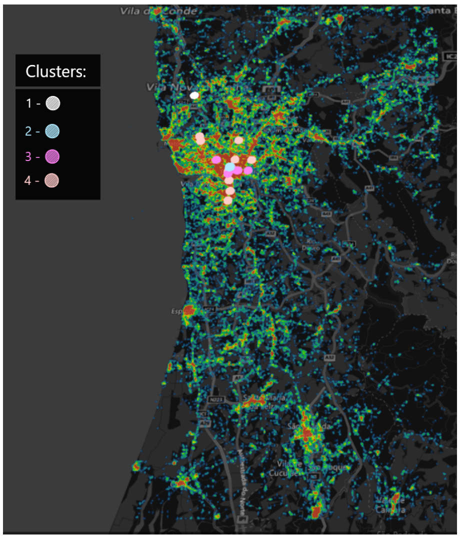
$$\begin{aligned} 40.833346 < Latitude < 41.358279 \\ -8.756861 < Longitude < -8.337497 \end{aligned} \tag{5.2}$$

After the loop ends, the types are converted into one of the 11 main categories mentioned in Section 3, and every POI within the city is stored with the following attributes:

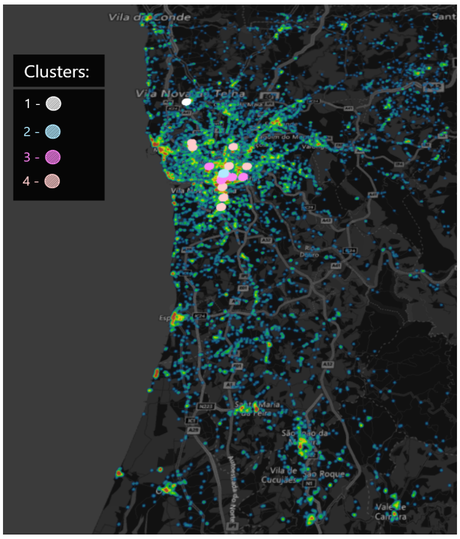
- **Name** <String>;
- **Category** <String>;
- **Latitude** <Float>;
- **Longitude** <Float>.

Statistics regarding all categories are shown in Table 5.6 while Figure 5.6 displays a heat-map with all points of interest for each category and the stations of the first 4 clusters defined in the spatial-temporal characterization. From this, we can see that:

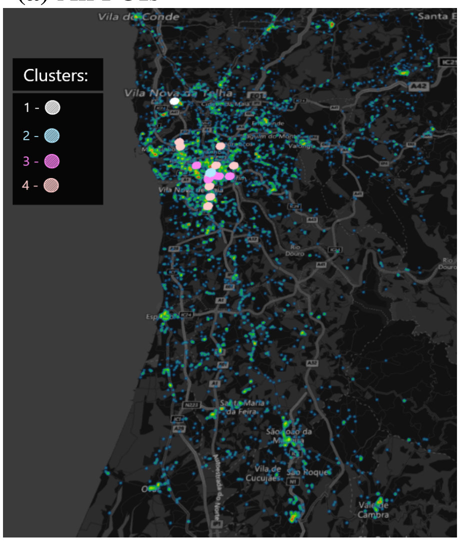
- Some categories' points of interest are concentrated in Porto's city center (e.g. Transportation and Lodging POIs);
- Some categories' points of interest are mainly concentrated in city centers (e.g. Service and Lifestyle and Beauty POIs);
- Some categories' points of interest are well spread across the Metropolitan Area of Porto (e.g. Government and Religion POIs);



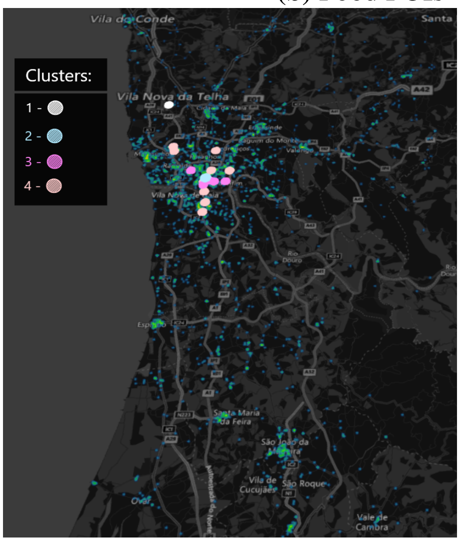
(a) All POIs



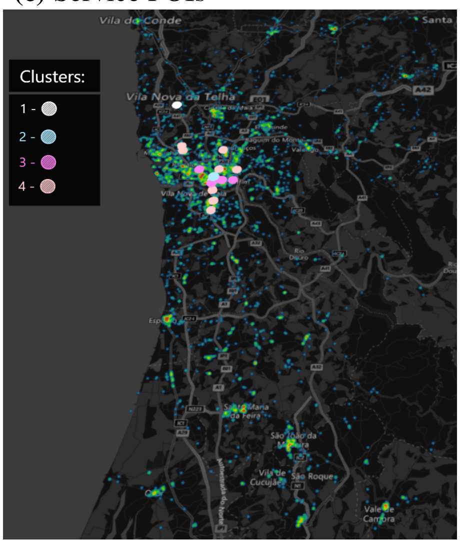
(b) Food POIs



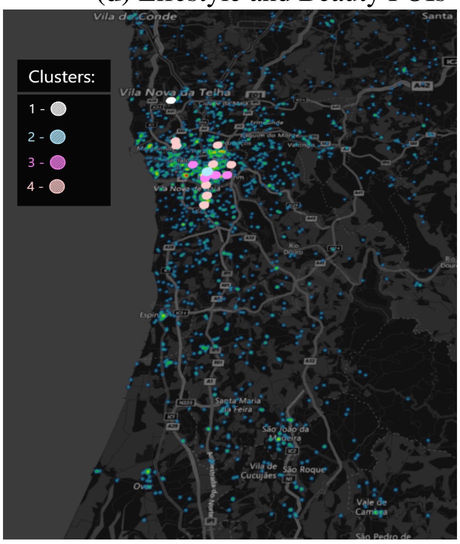
(c) Service POIs



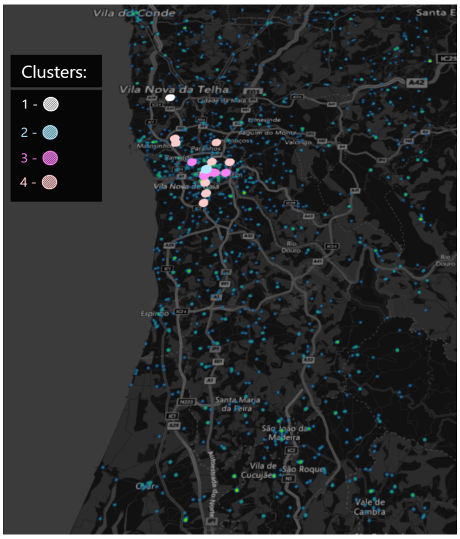
(d) Lifestyle and Beauty POIs



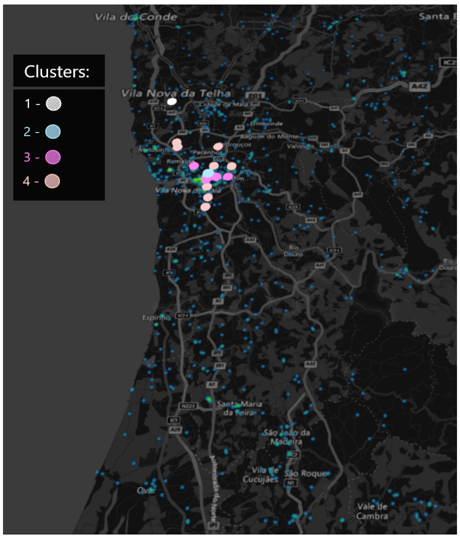
(e) Healthcare POIs



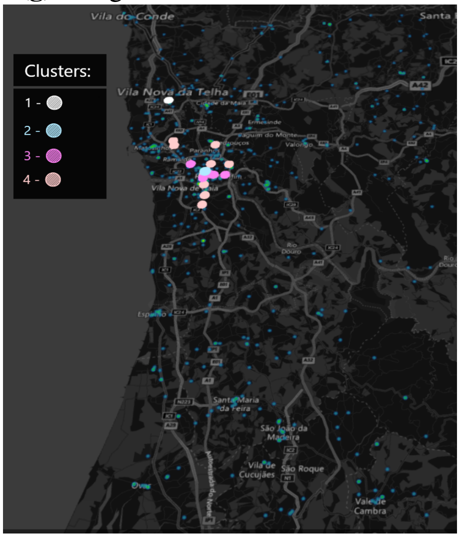
(f) EducationPOIs



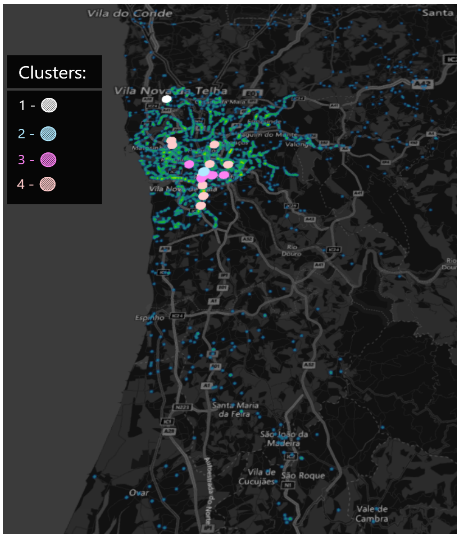
(g) Religion POIs



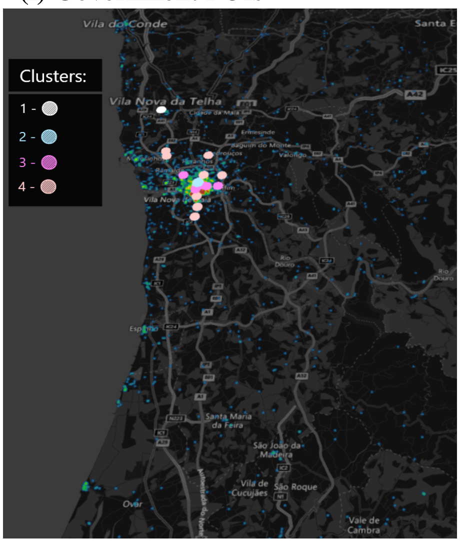
(h) Leisure and Culture POIs



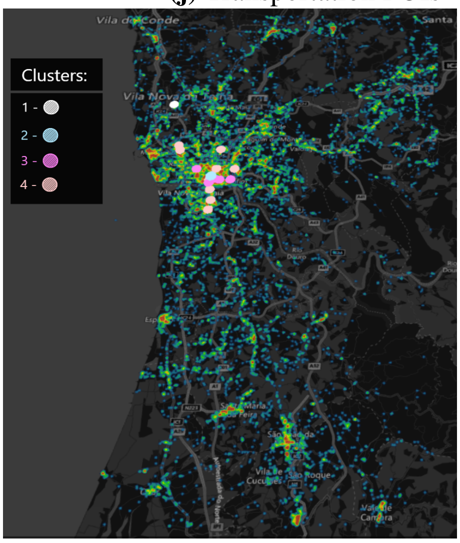
(i) Government POIs



(j) Transportation POIs



(k) Lodging POIs



(l) Shopping POIs

Figure 5.6: Map of Metropolitan Area of Porto with heat-map of Points of Interest and stations from Cluster 1-4

Table 5.6: Points of Interest' Categories Statistics

Category	**Nº of POI**
Food	7330
Service	4977
Lifestyle and Beauty	2397
Healthcare	3426
Education	1663
Religion	1314
Leisure and Culture	786
Government	503
Transportation	2037
Lodging	1818
Shopping	13511
All	39762

5.3.2 Pre-Processing

Firstly, we needed to obtain the coordinates of each station in the Metropolitan Area of Porto for our spatial characterization. To do such, we used another database, which contained, for each station, the following information:

- Station Name;
- Latitude;
- Longitude;

Using each station's latitude and longitude, we were able to calculate the number of points of interest, for each one of the 11 main categories, of the stations within clusters 1 through 4 originated in the spatial-temporal analysis. We selected these stations since they presented a higher amount of validations by users predicted the tourists.

In order to obtain the POIs for each station, we used a walking radius of 800 meters, which is the conventionally assumed threshold for walking to rail transports ([25]). Meaning that a POI is only counted as one of the POIs of a station if its distance to the station in question is less than 800 meters.

In the end, the method we used to collect the POIs for each station in the database can be seen in Table 5.7, where every column other than the first corresponds to the number of POIs for that category, in a radius of 800 meters to the station.

Table 5.7: Approach used while storing data for the Spatial characterization

Station	Food	Service	...	Lodging	Shopping
Station1	21	52	...	10	25
Station2	58	18	...	7	51
...					

5.3.3 K-Means

In order to be able to group the stations by the amount and category of points of interest within walking distance of them, we used the K-Means clustering algorithm.

But firstly, we calculated the optimal number of clusters and grouped all categories that are appealing to tourists and all categories intended for citizens. Each group consisted of the following categories:

- **Touristic Points Of Interest** - Food, Leisure and Culture, Transportation, Lodging and Shopping;
- **Citizens Points Of Interest** - Service, Lifestyle and Beauty, Healthcare, Education, Religion and Government.

After that process, the final data for the Points of Interest had the format shown in Table 5.8.

Table 5.8: Approach used while storing data for the Spatial characterization

Station	TouristicPOI	CitizenPOI
Station1	68	70
Station2	148	91
...		

5.3.4 Number of Clusters

To determine the optimal number of clusters, we chose to re-use the elbow method with distortion as our scoring metric.

From the elbow curve obtained (displayed in Figure 5.7), we opted to select three as the number of clusters for our k-means algorithm since it was the highest number of clusters with significantly lower distortion than the previous number.

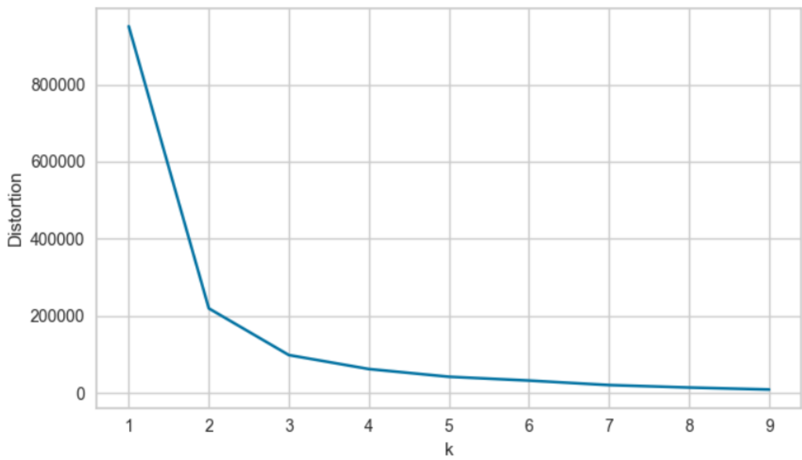


Figure 5.7: Elbow method result with distortion scores

5.3.5 Discussion Results

By executing the K-Means algorithm with 3 clusters, Figure 5.8 was generated with the following stations:

- **Blue Cluster**: Trindade, São Bento, Bolhão, Aliados;
- **Red Cluster**: 24 de Agosto, Campanhã, Estádio do Dragão, H. São João, Santo Ovídio, Srª da Hora, Sete Bicas;
- **Green Cluster**: Aeroporto, Casa da Música, João de Deus, Marquês, Jardim do Morro;

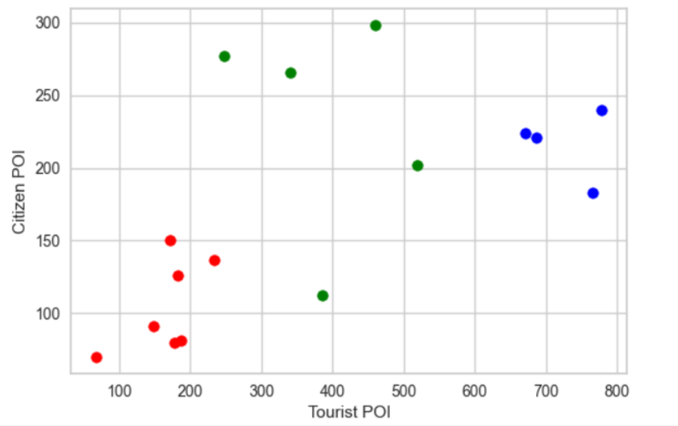


Figure 5.8: Plot of the clusters generated in Spatial Characterization

From these results, we were able to classify the clusters as follows:

- **Touristic Stations**: Stations in the Blue Cluster. These stations are classified as touristic due to the higher amount of Tourist POIs near them;
- **Non-Touristic Stations**: Stations in the Red Cluster. These stations have the lowest amount of POIs near them, making their location not appealing to tourists. However, this cluster also includes intermodality stations that tourists will also visit despite their location (e.g., "Campanha", "24 Agosto");
- **Semi-Touristic Stations**: Stations in the Green Cluster. In this cluster, these stations have a moderate amount of Tourist POIs near them but a varying amount of Citizen POIs. But, considering the stations divided into this cluster, we can observe that this cluster contains locations that may be attractive for some tourists, depending on their interests.

Finally, as we can see from Figure 5.9,

- The touristic stations are located in the city center, which is the most touristic place in Porto.
- The Non-Touristic Stations are well spread and not located in the most touristic areas;
- The Semi-Touristic stations can be near touristic localities but also in areas more occupied by residents.

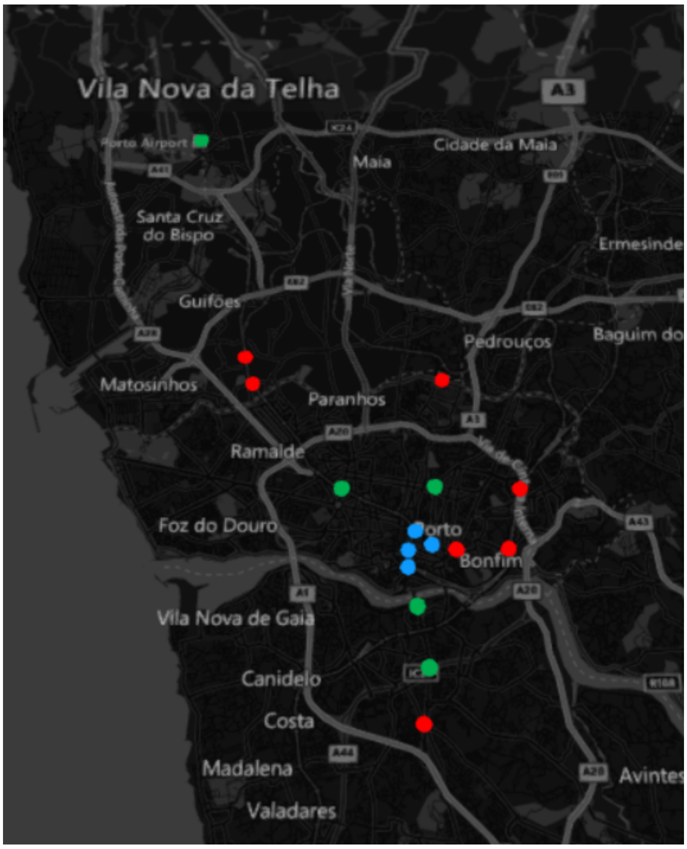


Figure 5.9: Map of Porto with the stations clustered in the Spatial Characterization

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Cities worldwide have started to adopt an Automated Fare Collection (AFC) system for their public transportation network, which opened new opportunities beyond fare collection. These systems generate a vast quantity of data regarding on-board transactions, which can then be used for research studies.

Over the years, there has been a massive increase in tourist activity, which derived into tough challenges in many cities that were not initially prepared to receive such an amount of tourists. To explore this growth and make tourism more sustainable, it would be vital to understand and identify movement patterns within the city for tourists. Since this case is evident for Porto, which has become an attractive destination for tourism with record-breaking numbers almost every year, the data gathered from the Automated Fare Collection system can be further explored for these non-frequent passengers.

Our main objective during our study was to be able to differentiate tourists from other users by the validations recorded in the travel cards. To do such, we used the different temporal patterns registered by Tour travel cards and Monthly Subscriptions cards and generated a scoring system that predicts that a travel card is being used by a tourist or not. From the results obtained, we can see that the scoring system can separate tourists from regular users with high accuracy but has difficulties identifying tourists from occasional users.

With the tourists identified, our next objective was to formalize three types of patterns for these users, with those patterns being Spatial Patterns, Temporal Patterns, and Spatial-Temporal Patterns. From these, we extracted valuable knowledge for tourists within the city so we could understand their mobility profiles and consequently, provide them a mobility service more tailored to their needs, while also taking into account the context the patterns are inserted in by relating them with data from the city's point of interest.

From the results of the patterns, we can see that the movement of tourists registered considerably distinct temporal patterns than non-tourists. While non-tourists have a higher percentage of validations during rush hours and weekdays, tourists had a higher percentage of validations during conventional tourism seasons (April to October).

As for the spatial-temporal distribution, we clustered the stations in the Metropolitan Area of Porto according to their daily distribution of validations from tourists and obtained similar results for most stations, with a lower percentage of validations during rush hours.

Finally, by counting the points of interest within walking distance of each station for each category, we were able to group and identify Touristic stations, Inhabitant stations, and Inconclusive stations to realize that all the touristic stations were at the most touristic spots of the city.

6.2 Future Work

The proposed method of identifying tourists is efficient in separating tourists from regular users. However, the distinction between tourists and occasional passengers still has room for improvement. So, future work could involve adding other features that help in distinguishing tourists from occasional citizens. One feature could be the classification of the station where the validation was performed, according to the type of points of interest nearby. These types could vary from Education, Health, Line Changing, Shopping, Eating, Etc.

With regard to the characterization of tourist mobility patterns, as this is not the focus of our study, this was not explored to the full. So, to build on this, we could use a different and more detailed approach regarding the temporal and spatial-temporal characterization in the future when generating the temporal profiles, such as a Trip Chaining Model. As spatial data was not fully available for bus trips, a regularly used means of transport, a future approach for spatial and spatial-temporal characterization would include these validations. Another factor that could also be used is predicting the destination of a trip rather than using the originating station.

Appendix A

Scoring System Results

Table A.1: Total results for TripsDayWeek variable of scoring system

	Tour1 & Tour3		**Monthly Subscription**	
	Total	**Percentage**	**Total**	**Percentage**
Monday	147695	13.19%	24063895	17.16%
Tuesday	152938	13.65%	24857188	17.72%
Wednesday	161706	14.44%	24466734	17.45%
Thursday	164850	14.72%	24116020	17.20%
Friday	177800	15.87%	24149853	17.22%
Saturday	169934	15.17%	11309510	8.06%
Sunday	145242	12.97%	7276293	5.19%

Table A.2: Total results for TripsHourWeekday variable of scoring system

	Tour1 & Tour3		**Monthly Subscription**	
	Total	**Percentage**	**Total**	**Percentage**
00:00-01:00	3204	0.40%	629045	0.50%
01:00-02:00	489	0.06%	112072	0.09%
02:00-03:00	219	0.03%	59066	0.05%
03:00-04:00	271	0.03%	46920	0.04%
04:00-05:00	267	0.03%	56559	0.04%
05:00-06:00	812	0.10%	320126	0.25%
06:00-07:00	3636	0.45%	3292246	2.60%
07:00-08:00	7186	0.89%	10640794	8.40%
08:00-09:00	19809	2.46%	13115568	10.35%
09:00-10:00	50614	6.29%	8560275	6.75%
10:00-11:00	72775	9.04%	6304899	4.97%
11:00-12:00	66525	8.26%	5710404	4.51%
12:00-13:00	58710	7.29%	6532012	5.15%
13:00-14:00	54764	6.80%	8480170	6.69%
14:00-15:00	55008	6.83%	7601077	6.00%
15:00-16:00	63504	7.89%	7465672	5.89%
16:00-17:00	64556	8.02%	9123168	7.20%
17:00-18:00	65753	8.17%	12157585	9.59%
18:00-19:00	62737	7.79%	11724285	9.25%
19:00-20:00	52666	6.54%	7165598	5.65%
20:00-21:00	38512	4.78%	3453847	2.72%
21:00-22:00	29706	3.69%	1682195	1.33%
22:00-23:00	21664	2.69%	1356181	1.07%
23:00-24:00	11602	1.44%	1158444	0.91%

Table A.3: Total results for TripsHourWeekend variable of scoring system

	Tour1 & Tour3		**Monthly Subscription**	
	Total	**Percentage**	**Total**	**Percentage**
00:00-01:00	2203	0.70%	362370	1.89%
01:00-02:00	393	0.12%	84655	0.44%
02:00-03:00	216	0.07%	48256	0.25%
03:00-04:00	193	0.06%	37597	0.20%
04:00-05:00	180	0.06%	31703	0.17%
05:00-06:00	527	0.17%	65408	0.34%
06:00-07:00	1497	0.47%	364273	1.90%
07:00-08:00	2863	0.91%	684453	3.56%
08:00-09:00	7278	2.31%	1045882	5.45%
09:00-10:00	19943	6.33%	1294028	6.74%
10:00-11:00	29554	9.38%	1336944	6.96%
11:00-12:00	27993	8.88%	1278369	6.66%
12:00-13:00	23147	7.34%	1119524	5.83%
13:00-14:00	21387	6.79%	1161744	6.05%
14:00-15:00	20365	6.46%	1437737	7.49%
15:00-16:00	23265	7.39%	1484693	7.73%
16:00-17:00	24188	7.67%	1421734	7.40%
17:00-18:00	25311	8.03%	1435386	7.47%
18:00-19:00	23366	7.41%	1372461	7.15%
19:00-20:00	20249	6.42%	1112196	5.79%
20:00-21:00	14957	4.75%	743084	3.87%
21:00-22:00	11494	3.65%	493221	2.57%
22:00-23:00	9417	2.99%	414558	2.16%
23:00-24:00	5175	1.64%	376907	1.96%

Table A.4: Total results for TripsMonth variable of scoring system

	Tour1 & Tour3		**Monthly Subscription**	
	Total	**Percentage**	**Total**	**Percentage**
January	36658	3.27%	9681008	6.90%
February	45098	4.03%	9452571	6.74%
March	68095	6.08%	10243702	7.30%
April	110304	9.85%	10544074	7.52%
May	117783	10.51%	13275558	9.47%
June	112614	10.05%	11170135	7.97%
July	120928	10.80%	12195381	8.70%
August	128363	11.46%	9747755	6.95%
September	126714	11.31%	12603314	8.99%
October	124855	11.15%	15047327	10.73%
November	74006	6.61%	13796652	9.84%
December	54747	4.89%	12482016	8.90%

Table A.5: Total results for NumberSequences variable of scoring system

	Tour1 & Tour3		**Monthly Subscription**	
	Total	**Percentage**	**Total**	**Percentage**
1	142247	92.47%	5970	6.90%
2	11552	7.51%	7097	6.74%
3	9	0.01%	8252	7.30%
4	4	0.00%	9702	7.52%
5	5	0.00%	10421	9.47%
6	4	0.00%	9826	7.97%
7	1	0.00%	8457	8.70%
8	1	0.00%	7988	6.95%
9	0	0.00%	7938	8.99%
10	0	0.00%	7779	10.73%
11	0	0.00%	7706	9.84%
12	0	0.00%	8395	8.90%
13	1	0.00%	8836	8.90%

Table A.6: Total results for NumberDaysPerSequence variable of scoring system

	Tour1 & Tour3		**Monthly Subscription**	
	Total	**Percentage**	**Total**	**Percentage**
1	65708	39.709%	3454989	32.158%
2	37307	22.546%	1907247	17.752%
3	34339	20.752%	1237160	11.515%
4	28107	16.986%	989763	9.212%
5	7	0.004%	2010234	18.711%
6	2	0.001%	930587	8.662%
7	2	0.001%	140990	1.312%
8	0	0%	72862	0.678%

Table A.7: Total results for NumberTripsPerSequence variable of scoring system

	Tour1 & Tour3		Monthly Subscription	
	Total	**Percentage**	**Total**	**Percentage**
1	18716	11.311%	765667	6.877%
2	18796	11.359%	1364096	12.252%
3	16817	10.163%	797657	7.165%
4	16272	9.834%	1012703	9.096%
5	13985	8.452%	619756	5.567%
6	12023	7.266%	636250	5.715%
7	10266	6.204%	462561	4.155%
8	9146	5.527%	531485	4.774%
9	7700	4.653%	427593	3.841%
10	6896	4.167%	520414	4.674%
11	6034	3.647%	364939	3.278%
12	5287	3.195%	373626	3.356%
13	4487	2.712%	285819	2.567%
14	3781	2.285%	269133	2.417%
15	3220	1.946%	243107	2.184%
16	2584	1.562%	236943	2.128%
17	2055	1.242%	199586	1.793%
18	1668	1.008%	196357	1.764%
19	1336	0.807%	179010	1.608%
20	1080	0.653%	205790	1.848%
21	816	0.493%	152329	1.368%
22	619	0.374%	135683	1.219%
23	477	0.288%	116674	1.048%
24	396	0.239%	108008	0.97%
25	272	0.164%	93041	0.836%
26	170	0.103%	80726	0.725%
27	145	0.088%	71057	0.638%
28	106	0.064%	63705	0.572%
29	68	0.041%	56760	0.51%
30	52	0.031%	53512	0.481%
31	32	0.019%	45542	0.409%
32	36	0.022%	39969	0.359%
33	14	0.008%	35483	0.319%
34	22	0.013%	32140	0.289%
35	12	0.007%	28689	0.258%
36	8	0.005%	26116	0.235%
37	10	0.006%	23363	0.21%
38	8	0.005%	21037	0.189%
39	2	0.001%	19337	0.174%
40	6	0.004%	17961	0.161%
41	10	0.006%	16236	0.146%
42	10	0.006%	15019	0.135%
43	5	0.003%	13998	0.126%
44	1	0.001%	13256	0.119%
45	4	0.002%	12165	0.109%
46	1	0.001%	11706	0.105%
47	2	0.001%	10912	0.098%
48	1	0.001%	10318	0.093%
49	3	0.002%	9703	0.087%
50	2	0.001%	8933	0.08%
51	2	0.001%	8465	0.076%
52	1	0.001%	8096	0.073%
53	3	0.002%	7540	0.068%
54	0	0%	7470	0.067%
55	1	0.001%	6938	0.062%
56	1	0.001%	6728	0.06%
57	0	0%	6268	0.056%
58	0	0%	6150	0.055%
59	1	0.001%	5738	0.052%
60	1	0.001%	5383	0.048%
61	0	0%	5242	0.047%
62	0	0%	5041	0.045%
63	0	0%	4886	0.044%
64	1	0.001%	4708	0.042%
65	0	0%	4490	0.04%
66	1	0.001%	4306	0.039%

Table A.8: Total results for HasCycle variable of scoring system

	Tour1 & Tour3		**Monthly Subscription**	
	Total	**Percentage**	**Total**	**Percentage**
Yes	3	0.00%	139380	37.25%
No	153821	100%	234710	62.74%

Appendix B

Pattern Formalization Results

Table B.1: Temporal results for Weekly Distribution

	Tourist Cards		Non-Tourist Cards	
	Total	**Percentage**	**Total**	**Percentage**
Monday	927074	13.62%	12828773	16.99%
Tuesday	949430	13.95%	13135353	17.4%
Wednesday	924564	13.58%	12971822	17.18%
Thursday	946442	13.9%	12772603	16.92%
Friday	1031892	15.16%	12785043	16.93%
Saturday	1072441	15.76%	6505655	8.62%
Sunday	954779	14.03%	4499547	5.96%

Table B.2: Temporal results for Daily distribution on Weekdays and Weekend

	Tourist Cards		Non-Tourist Cards	
	Total	**Percentage**	**Total**	**Percentage**
$H_0 0$	40780	0.85%	405495	0.63%
$H_0 1$	13400	0.28%	47489	0.07%
$H_0 2$	6038	0.13%	11475	0.02%
$H_0 3$	4490	0.09%	10463	0.02%
$H_0 4$	3820	0.08%	16107	0.02%
$H_0 5$	10696	0.22%	131562	0.2%
$H_0 6$	43054	0.9%	1269661	1.97%
$H_0 7$	82565	1.73%	4851738	7.52%
$H_0 8$	168904	3.53%	6892830	10.69%
$H_0 9$	305744	6.4%	4298394	6.66%
$H_0 10$	383598	8.03%	2946387	4.57%
$H_0 11$	372529	7.79%	2665149	4.13%
$H_0 12$	338122	7.07%	3197368	4.96%
$H_0 13$	331809	6.94%	4085452	6.33%
$H_0 14$	335310	7.02%	3662071	5.68%
$H_0 15$	372775	7.8%	3652979	5.66%
$H_0 16$	374680	7.84%	4477682	6.94%
$H_0 17$	380374	7.96%	6406289	9.93%
$H_0 18$	355841	7.45%	6647105	10.31%
$H_0 19$	286338	5.99%	3976358	6.17%
$H_0 20$	195048	4.08%	1969473	3.05%
$H_0 21$	150758	3.15%	1111835	1.72%
$H_0 22$	131680	2.76%	955446	1.48%
$H_0 23$	91049	1.91%	804786	1.25%
$H_1 0$	28561	1.41%	262538	2.39%
$H_1 1$	5078	0.25%	41650	0.38%
$H_1 2$	1537	0.08%	11069	0.1%
$H_1 3$	889	0.04%	6552	0.06%
$H_1 4$	779	0.04%	6345	0.06%
$H_1 5$	5567	0.27%	35767	0.33%
$H_1 6$	19299	0.95%	167958	1.53%
$H_1 7$	28029	1.38%	311354	2.83%
$H_1 8$	54509	2.69%	497945	4.52%
$H_1 9$	110716	5.46%	615719	5.59%
$H_1 10$	141652	6.99%	628718	5.71%
$H_1 11$	152666	7.53%	631513	5.74%
$H_1 12$	135316	6.67%	607390	5.52%
$H_1 13$	130309	6.43%	640713	5.82%
$H_1 14$	128186	6.32%	800177	7.27%
$H_1 15$	154002	7.6%	864024	7.85%
$H_1 16$	162844	8.03%	847085	7.7%
$H_1 17$	169488	8.36%	878507	7.98%
$H_1 18$	159609	7.87%	872319	7.93%
$H_1 19$	141971	7.00%	760812	6.91%
$H_1 20$	97717	4.82%	527351	4.79%
$H_1 21$	76583	3.78%	370661	3.37%
$H_1 22$	72352	3.57%	329765	3.00%
$H_1 23$	49561	2.44%	289270	2.63%

Table B.3: Temporal results for Monthly Distribution

	Tourist Cards		**Non-Tourist Cards**	
	Total	**Percentage**	**Total**	**Percentage**
January	396715	5.83%	5347438	7.08%
February	351233	5.16%	5264702	6.97%
March	467330	6.87%	5763607	7.63%
April	570849	8.39%	7591861	10.06%
May	641344	9.42%	6867704	9.1%
June	675808	9.93%	5731357	7.59%
July	694714	10.21%	6131160	8.12%
August	772054	11.34%	4998256	6.62%
September	686009	10.08%	6460904	8.56%
October	612249	8.99%	7704142	10.2%
November	418393	6.15%	7055016	9.34%
December	519924	7.64%	6582649	8.72%

Table B.4: Spatial-Temporal results for Daily distribution on Weekdays and Weekend (Clusters 1,2,3)

	Cluster 1	**Cluster 2**	**Cluster 3**				
	Aeroporto	**Trindade**	**São Bento**	**Campanhã**	**Casa Música**	**Bolhão**	**24 Agosto**
$H_0 0$	7171	3668	2660	2644	2007	1986	1465
$H_0 1$	2474	1151	839	843	655	673	459
$H_0 2$	1097	496	395	366	313	312	209
$H_0 3$	776	410	303	271	236	212	160
$H_0 4$	665	362	268	236	176	207	114
$H_0 5$	1882	930	707	691	558	521	397
$H_0 6$	7611	3876	2855	2820	2090	2093	1452
$H_0 7$	14811	7419	5276	5140	4135	4145	2715
$H_0 8$	30518	15268	10659	10713	8240	8101	5821
$H_0 9$	55019	27097	19547	19511	15264	14512	10564
$H_0 10$	69149	34325	24670	24432	18815	18072	12866
$H_0 11$	67309	33139	23911	23713	18090	17398	12608
$H_0 12$	60389	30270	21903	21525	16829	15849	11434
$H_0 13$	59750	29472	21188	20972	16551	15734	11220
$H_0 14$	59745	30294	21562	21537	16270	15887	11292
$H_0 15$	67503	33328	23817	23332	18393	17477	12571
$H_0 16$	66979	33454	24161	23951	18656	17697	12755
$H_0 17$	67810	33985	24616	23926	18659	17936	12928
$H_0 18$	63733	31695	22884	22514	17566	16955	11909
$H_0 19$	51160	25632	18272	18117	14292	13814	9641
$H_0 20$	35374	17459	12467	12239	9549	9178	6660
$H_0 21$	27217	13313	9540	9675	7304	7297	4997
$H_0 22$	23548	11810	8280	8517	6552	6219	4620
$H_0 23$	16217	8202	5745	5708	4391	4267	3118
$H_1 0$	5076	2447	1808	1732	1473	1293	1023
$H_1 1$	920	442	316	321	280	246	165
$H_1 2$	262	147	109	77	79	74	39
$H_1 3$	145	76	55	59	52	38	29
$H_1 4$	142	64	46	46	30	40	27
$H_1 5$	1003	455	351	354	282	269	166
$H_1 6$	3521	1739	1262	1253	926	919	669
$H_1 7$	5133	2506	1764	1784	1416	1313	987
$H_1 8$	9705	4927	3467	3542	2747	2581	1871
$H_1 9$	19919	9844	6883	7012	5543	5224	3795
$H_1 10$	25851	12770	9172	9046	6805	6724	4843
$H_1 11$	27203	13650	9760	9674	7539	7129	5136
$H_1 12$	24286	12094	8606	8685	6494	6490	4559
$H_1 13$	23375	11732	8376	8324	6583	6150	4451
$H_1 14$	23074	11505	8373	8108	6356	6061	4410
$H_1 15$	27606	13689	9930	9761	7669	7333	5237
$H_1 16$	29307	14543	10533	10204	7998	7732	5603
$H_1 17$	30446	14977	10835	10805	8566	8311	5716
$H_1 18$	28488	14022	10389	10135	7942	7607	5449
$H_1 19$	25221	12643	9130	9091	6886	6967	4807
$H_1 20$	17448	8638	6252	6307	4758	4694	3279
$H_1 21$	13737	6794	4965	5043	3712	3596	2571
$H_1 22$	13085	6322	4527	4589	3549	3562	2422
$H_1 23$	8724	4347	3166	3189	2516	2354	1662

Table B.5: Spatial-Temporal results for Daily distribution on Weekdays and Weekend (Cluster 4)

	Cluster 4								
	João de Deus	**Estádio Dragão**	**H. São João**	**Aliados**	**Santo Oví-dio**	**Sr Hora**	**Marquês**	**Sete Bicas**	**Jardim Morro**
$H_0$0	889	836	850	831	811	706	665	675	629
$H_0$1	313	252	277	239	229	256	225	212	213
$H_0$2	136	102	121	122	120	115	114	99	92
$H_0$3	96	134	126	67	93	72	75	85	63
$H_0$4	73	88	84	73	63	75	61	65	78
$H_0$5	236	225	260	169	188	196	206	188	149
$H_0$6	967	858	880	857	854	772	709	758	626
$H_0$7	1782	1772	1650	1703	1493	1616	1326	1401	1182
$H_0$8	3659	3526	3408	3507	3245	2936	2868	2738	2537
$H_0$9	6773	6381	6237	6336	5552	5478	5108	5329	4527
$H_0$10	8479	8137	7854	7764	7163	7057	6456	6243	5786
$H_0$11	8275	8055	7689	7487	6945	6837	6238	6108	5724
$H_0$12	7559	6954	7001	7014	6257	6084	6001	5677	4999
$H_0$13	7116	6990	6894	6833	6281	5974	5624	5453	4934
$H_0$14	7285	7044	6825	6945	6315	5899	5671	5504	5001
$H_0$15	8119	7668	7641	7737	6974	6771	6356	6169	5516
$H_0$16	8184	7790	7675	7662	7172	6782	6279	6230	5612
$H_0$17	8269	8042	8017	7581	7322	6905	6362	6306	5739
$H_0$18	7754	7331	7480	7357	6570	6344	6047	6051	5365
$H_0$19	6291	6174	5968	5751	5352	5140	4889	4644	4183
$H_0$20	4248	4002	4042	4007	3725	3465	3329	3191	2888
$H_0$21	3270	3130	3098	3135	2808	2695	2613	2432	2307
$H_0$22	2945	2841	2651	2672	2440	2202	2302	2192	1994
$H_0$23	1998	2029	1914	1855	1768	1622	1644	1422	1370
$H_1$0	611	612	581	641	568	471	502	462	432
$H_1$1	98	103	113	124	104	83	78	75	73
$H_1$2	34	39	29	35	32	18	26	29	21
$H_1$3	22	20	25	20	16	17	21	9	10
$H_1$4	18	20	18	20	13	16	16	7	11
$H_1$5	142	123	108	103	104	103	112	86	77
$H_1$6	437	409	409	391	376	289	301	314	283
$H_1$7	563	568	633	590	506	504	498	462	419
$H_1$8	1237	1192	1074	1085	1069	975	996	845	791
$H_1$9	2468	2394	2397	2237	2059	1993	1914	1875	1560
$H_1$10	3148	2975	2853	2934	2658	2473	2353	2390	2166
$H_1$11	3492	3243	3180	3246	2834	2746	2557	2539	2169
$H_1$12	2917	2902	2884	2915	2533	2471	2229	2314	1956
$H_1$13	2884	2772	2652	2709	2377	2320	2246	2167	1864
$H_1$14	2826	2721	2675	2585	2442	2267	2095	2094	1917
$H_1$15	3450	3201	3103	3157	2911	2821	2564	2597	2241
$H_1$16	3691	3291	3351	3292	3065	2908	2762	2711	2497
$H_1$17	3577	3605	3557	3502	3211	3118	2779	2651	2486
$H_1$18	3606	3416	3296	3208	2981	2876	2681	2642	2328
$H_1$19	3101	2915	3012	2785	2682	2543	2426	2358	2112
$H_1$20	2151	2095	2084	2022	1913	1784	1654	1592	1427
$H_1$21	1734	1638	1476	1577	1458	1298	1282	1299	1095
$H_1$22	1624	1513	1459	1482	1353	1300	1250	1173	1087
$H_1$23	1108	1019	1027	1075	948	845	797	797	756

References

[1] Bruno Agard, Catherine Morency, and Martin Trépanier. Mining public transport user behaviour from smart card data. 01 2006.

[2] Daniel Albalate and Germa Bel. What shapes local public transportation in europe? economics, mobility, institutions, and geography. *Transportation Research Part E: Logistics and Transportation Review*, 46:775–790, 09 2010.

[3] Atizaz Ali, Jooyoung Kim, and Seungjae Lee. Travel behavior analysis using smart card data. *KSCE Journal of Civil Engineering*, 20, 07 2015.

[4] Ana Azevedo and Manuel Santos. Kdd, semma and crisp-dm: A parallel overview. pages 182–185, 01 2008.

[5] James Barry, Robert Newhouser, Adam Rahbee, and Shermeen Sayeda. Origin and destination estimation in new york city with automated fare system data. *Transportation Research Record*, 1817:183–187, 01 2002.

[6] Philip Blythe. Improving public transport ticketing through smart cards. *Proceedings of the ICE - Municipal Engineer*, 157:47–54, 01 2004.

[7] Artem Chakirov and Alexander Erath. Use of public transport smart card fare payment data for travel behaviour analysis in singapore. 2011.

[8] Etienne Côme and Latifa Oukhellou. Model-based count series clustering for bike sharing system usage mining: A case study with the vélib' system of paris. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5, 10 2014.

[9] Ding Du, Alan Lew, and Pin Ng. Tourism and economic growth. *Journal of Travel Research*, 55:454–464, 12 2016.

[10] M. K. El Mahrsi, E. Côme, L. Oukhellou, and M. Verleysen. Clustering smart card data for urban mobility analysis. *IEEE Transactions on Intelligent Transportation Systems*, 18(3):712–728, 2017.

[11] Marta Ferreira, Vera Costa, Teresa Dias, and João Falcão e Cunha. Understanding commercial synergies between public transport and services located around public transport stations. *Transportation Research Procedia*, 27:125–132, 01 2017.

[12] Marta Ferreira, Teresa Dias, and João Falcão e Cunha. *With Whom Transport Operators Should Partner? An Urban Mobility and Services Geolocation Data Analysis: Second EAI International Conference, INTSYS 2018, Guimarães, Portugal, November 21–23, 2018, Proceedings*, pages 133–143. 01 2019.

[13] Zuoxian Gan, Min Yang Yang, Tao Feng, and Harry Timmermans. Understanding urban mobility patterns from a spatiotemporal perspective: daily ridership profiles of metro stations. *Transportation*, 47, 02 2020.

[14] Inês Gusman, Chamusca, Fernandes, and Jorge Ricardo Pinto. Culture and tourism in porto city centre: Conflicts and (im)possible solutions. *Sustainability*, 11:5701, 10 2019.

[15] Aaron Gutiérrez, Antoni Domènech, Benito Zaragozí, and Daniel Miravet. Profiling tourists' use of public transport through smart travel card data. *Journal of Transport Geography*, 88:102820, 10 2020.

[16] Stefan Gössling, Daniel Scott, and Colin Hall. Global trends in length of stay: implications for destination management and climate change. *Journal of Sustainable Tourism*, 26:1–15, 12 2018.

[17] Yu Lu, Huayu Wu, Liu Xin, Penghe Chen, and Jiyong Zhang. Toursense: A framework for tourist identification and analytics using transport data. *IEEE Transactions on Knowledge and Data Engineering*, PP:1–1, 01 2019.

[18] Xiaolei Ma, Yao-Jan Wu, Y. Wang, Feng Chen, and Jianfeng Liu. Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies*, 36:1–12, 11 2013.

[19] Marcela Munizaga and Carolina Palma. Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smartcard data from santiago, chile. *Transportation Research Part C-Emerging Technologies*, 24:9–18, 10 2012.

[20] D. Namiot and M. Schneps-Schneppe. A survey of smart cards data mining. In *AIST*, 2017.

[21] Kamal Nigam, Andrew Mccallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39:103–134, 05 2000.

[22] Diogo Rafael Pinto Nogueira. Agile data mining: uma metodologia ágil para o desenvolvimento de projetos de data mining. 2014.

[23] Meisy Ortega-Tong. Classification of london's public transport users using smart card data. 12 2013.

[24] Fernando Plumed, Lidia Contreras-Ochando, Cèsar Ferri, Jose Orallo, Meelis Kull, Nicolas Lachiche, M. Ramírez-Quintana, and Peter Flach. Crisp-dm twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*, PP:1–1, 12 2019.

[25] Dennis van Soest, Miles Tight, and Chris Rogers. Exploring the distances people walk to access public transport. *Transport Reviews*, 40:1–23, 02 2019.

[26] Mingqiang Xue, Huayu Wu, Wei Chen, and Gin Goh. Identifying tourists from public transport commuters. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 08 2014.

[27] Jinhua Zhao, Adam Rahbee, and Nigel Wilson. Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. *Comp.-Aided Civil and Infrastruct. Engineering*, 22:376–387, 07 2007.