FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Understanding mobility profiles of non-frequent public transport passengers through data extraction

Diogo André Fernandes Coelho

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Marta Campos Ferreira

Second Supervisor: Terese Galvão Dias

July 28, 2020

Understanding mobility profiles of non-frequent public transport passengers through data extraction

Diogo André Fernandes Coelho

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: Prof. Ana Paiva
External Examiner: Prof. Marco Veloso
Supervisor: Prof. Marta Campos Ferreira

July 28, 2020

Abstract

The public transport is an important resource for people in a city, even more, important in big cities. Important to minimize the utilization of individual transport and to give the possibility of everyone reaching any point in a city. Nowadays, most of the public transport networks utilize automated fare collection (AFC) systems to automate the ticketing system of some or several public transport modes. Such systems produce a large amount of very detailed data regarding on-board transactions. A lot of information can be extracted about frequent urban transport passengers, however little is known about occasional passengers (travel patterns, activities performed and travel behaviour). The purpose of this work is to understand the mobility patterns of non-frequent public transport passengers through data extraction. We use the city of Porto, Portugal, as an illustrative example. The electronic ticketing system used in Metropolitan Area of Porto (MAP) is an entry-only AFC system with a distance-based fare structure. The fare media are contactless travel cards, called Andante, which can be used on buses from several operators, as well as on metro and railways. There are two types of Andante travel cards: occasional tickets (OT) and monthly subscription (MS). Usually, MS is used by locals living in the city who use public transport regularly. OT is mostly used by occasional passengers and tourists. The data used in this work is the set of transaction records in 2013 from Sociedade de Transportes Colectivos do Porto (STCP) and metro of Porto, the public bus and metro operators, system respectively. The main objective of this study was to find patterns of non-frequent passengers. To achieve this purpose we defined three different types of patterns that could occur. Temporal patterns, the spatial patterns and last the spatial-temporal patterns, that is a conjunction of the previous two. First, the temporal patterns are where we profiled the temporal usage of the transport network of passengers. This consisted of processing the temporal data of the non-frequent passenger's data with GMM to create the temporal profiles for days and weeks. We found some patterns, for example, morning usage of the network, afternoon usage of the network and the combination of both. Also, we looked for the number of new non-frequent users of the network to understand the flow of new tickets each day throughout a year. Second, for the spatial patterns, we located what are the areas of the transport network that are more used by the non-frequent passengers. Finally, the Spatio-temporal patterns consisted of profiling the usage of the transport network in both time and space. We found that the results are almost equal throughout the year. In addition to these findings, we also looked for trip-chain in the non-frequent passengers. For this, we made trip-chain of zones and not trip-chain of stations. This way we discovered that the non-frequent passengers of the transport network have some trip-chain patterns. We found some pairs of trip-chain zones that are a pattern throughout the year.

Keywords: Data extraction, Public Transport, Mobility Patterns, AFC System

Acknowledgements

I would first like to thank all the members of my family that help me throughout these years. In special to my parents and girlfriend that were always there for me and always understood some of my crazy habits of work and, also, to my uncles that made possible for me to study at Porto and made me feel at home in every single day!

This journey would not be the same without friends. For all of them that made this journey enjoyable a big thanks and a wish of luck to all.

I would also like to thank my supervisors Marta Campos Ferreira and Teresa Galvão that were always there to help me with all the thesis problems and guide me to where to go next. I am very thankful that you let me be part of this project.

Diogo Coelho

Contents

List of Figures

List of Tables

Abbreviations

AFC	Automated Fare Collection
DBSCAN	Density-based spatial clustering of applications with noise
GMM	Gaussian Mixture Models
AIC	Akaike information criterion
BIC	Bayesian information criterion
MAP	Metropolitan Area of Porto
STCP	Sociedade de Transportes Colectivos do Porto

Chapter 1

Introduction

1.1 Problem

The pervasive adoption of Automated Fare Collection (AFC) systems by transport operators worldwide broadens the range of new possibilities beyond fare collection. Such systems produce a large amount of very detailed data regarding on-board transactions. A lot of information can be extracted about frequent urban transport passengers, however little is known about occasional passengers (travel patterns, activities performed and travel behaviour). This thesis aims to discover the mobility patterns of non-frequent users of public transport in the city of Porto. The metropolitan area of Porto(MAP) is composed of 17 municipalities, whereas one of these municipalities is the city of Porto. This city is the second biggest urban area of Portugal. In the MAP there is an entity called *Transportes Intermodais do Porto* (TIP) that was constituted by Metro do Porto, SA, pela Sociedade de Transportes Colectivos do Porto (STCP), SA e CP- Comboios de Portugal at 20 of December of 2002, to promote the implementation of inter-modality in public transport [16]. This entity created the ANDANTE system. The ANDANTE is an Automated Fare Collection (AFC) system that handles all the ticketing and fares of the transport operators operating in MAP and that is part of the Sistema Intermodal Andante(SIA) that are in it.

Enterprises always want to be the most efficient possible, not only to maximize their revenue but also to have satisfied their customer's needs. Regarding transport operators, their efficiency can be maximized in different aspects. One of these main aspects is making the transport network appropriate for their passengers/customers needs.

Nowadays, almost all transport operator utilizes an AFC system. This system is normally adopted because it eases the fare collection process and makes it more efficient. Such systems produce a large amount of detailed data regarding the on-board transactions of the transport. This information can be used to adapt/change the transport network to the passenger's needs. There is a lot of information that can be retrieved and has been retrieved from this large amount of data about frequent urban transport passengers. However, there is much less information about occasional passengers. There are less to none analysis about non-frequent passengers when compared to frequent passengers.

1.2 Motivation and Goals

The goal of this work is to extract knowledge from the AFC data to understand mobility profiles of non-frequent public transport passengers in different terms. In our case, the terms are Temporal Patterns, Spatial Patterns and Spatio-Temporal Patterns. This way we can discover patterns of utilization of non-frequent passengers data and with that make more suitable offers. If we find that non-frequent passengers have patterns of utilization in the transport network that means they also represent patterns of the transport network as the frequent passenger's data.

1.3 Structure of the Dissertation

In addition to the introduction, this dissertation has five more chapters. In Chapter 2 the State of the Art is presented. In Chapter 3 we talk about the problem and our approach. Chapter 4 describes all the processes and techniques done to retrieve knowledge from non-frequent passengers data. In Chapter 5 the achieved results are presented and discussed the achieved results. Finally, Chapter 6 presents the conclusion and future work.

Chapter 2

Literature Review

2.1 INTRODUCTION

In this context, the main aim of this study is to understand the mobility patterns of occasional passengers. To achieve this, data from the public transport of the city of Porto, Portugal, will be used. This dataset consists of the total number of validations made by public transport passengers at STCP and Metro of Porto during the 2013 year. To explore this large dataset, data mining techniques will be applied.

This Chapter2 is organized in the following way. First, an overview of Data Mining techniques and processes with a focus on Cross Industry Standard Process for Data Mining(CRISP-DM). The focus on the CRISP-DM process helps to apprehend where each data mining technique is realized. The third section is an overview of the Automated Fare Collection system and where each study fits on the different types of systems. Then, in the fourth section, it described the different cases where the estimation of an OD-Matrix was the focus since many studies do this approach in this type of data. The fifth section is about the Clustering approaches to find the passengers patterns of utilization.

2.2 Data Mining

In this section, we will describe and show some examples of techniques and processes that data mining provides to help us work on this objective of getting valuable information from the raw public transport data.

Data mining is the process of discovering hidden, valid and potentially useful patterns in datasets[17]. Different methodologies could be used, for example, KDD(Knowledge Discovery in Databases), SEMMA or even CRISP-DM. The methodology that we will try to use is CRISP-DM.

This methodology is a non-stop process as Fig. 2.1 shows. The target is always to fulfil the business objectives, so this process is always "evolving" their results to better suit the business objectives.

CRISP-DM is a methodology that is composed of 6 phases that are presented below.

Figure 2.1: CRISP-DM Methodology process

- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment

All these phases have their importance to the result and they are detailed below.

2.2.1 Business Understanding

The business understanding is the phase where we get familiarity with the problem that we need to solve. In this phase, some tasks should be performed. These tasks are:

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

2.2.2 Data Understanding

Data understanding is the phase where we try to collect and understand the data. The use of statistics in this phase could be useful to better describe the data. The tasks that we should consider in this phase are:

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

2.2.3 Data Preparation

The data preparation is the phase where we prepare the data to model it. This process consists of the following tasks:

- Select Data;
- Clean Data;
- Construct Data;
- Integrate Data;
- Format Data;

2.2.4 Modelling

In the modelling phase, the model is created. In this phase, we choose the algorithm that is appropriate to run on a certain type of data or the most appropriate for a certain objective. The tuning of the model parameters is also done in this phase. The tasks that should be considered in this phase are:

- Select the modelling technique;
- Generate Test Design;
- Build Model;
- Assess Model

This phase has different techniques and some of them are:

- Anomaly detection;
- Clustering;

- Classification;
- Regression;

Each one of these type of techniques has its importance to the result and are detailed below.

2.2.4.1 Anomaly detection

Anomaly detection is a technique that helps to identify unexpected data, as outliers. This can be useful to minimize their impact or to investigate what is the origin of this unexpected data.

2.2.4.2 Regression

Regression techniques are used to predict some variable from the relationship between some existing variables. Regression will not be used in this study since the objective is to find patterns in data and not to make predictions.

2.2.4.3 Classification

Classification is used to classify data in different classes. This type of technique is relevant to retrieve significant information from data, and some existent algorithms help to classify the data. Some of these algorithms are Linear Classifiers and Decision Trees.

Linear Classifiers These algorithms create a linear decision surface to classify the data, by other words these algorithms create a line that divides data into different classes. Figure 2.2 is an example of how the data is divided by these algorithms.

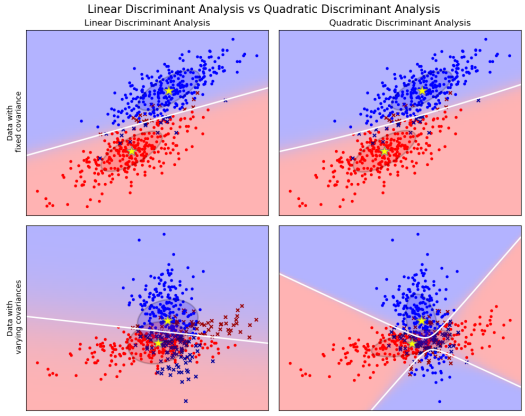


Figure 2.2: Examples of Linear Classifiers(LDA and QDA)− Reproduce from Scikit-learn [19]

Decision Trees Decision Tree is an algorithm that better shows the division of the data into classes since the reason for a division, normally, appears on the ramification. The end of each ramification is one class. Figure 2.3 represents the classes after running a decision tree on a certain dataset. At the beginning of each ramification is the cause of the division into ramifications.

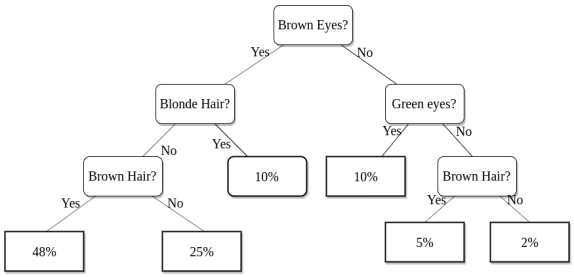


Figure 2.3: Example of a Decision Tree from Data with person characteristics

2.2.4.4 Clustering

Clustering is a technique that organizes data into groups. Normally, this technique finds the natural grouping between the elements of a given dataset by their variables. There are some existent algorithms that are used for this purpose, for example, K-means and DBSCAN.

K-Means K-mean is a clustering algorithm that needs the definition of the number of clusters by the user. The algorithm creates the number of centroids corresponding to the number of clusters inserted by the user and creates the clusters based on the distance of the existent data to those centroids. Figure 2.4 shows the centroids and the data corresponding to that centroid cluster.

DBSCAN DBSCAN is a density-based algorithm. By other words the clusters created in this algorithm result from how close certain parts of the data are. This algorithm needs the definition of the number of points in a neighborhood to consider it a cluster and the value of the distance that will outline points. These points create a cluster if they are inside the outline of each other. Figure 2.5 is a representation of the DBSCAN algorithm.

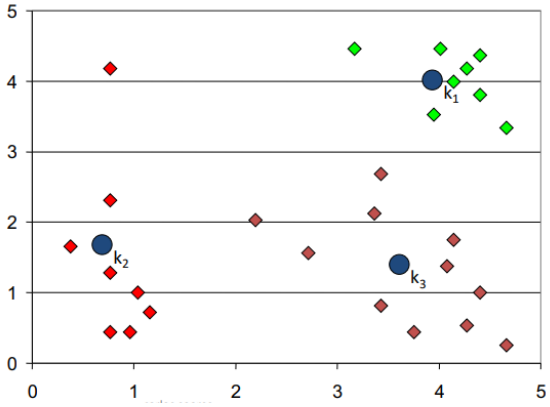


Figure 2.4: Example of K-means algorithm

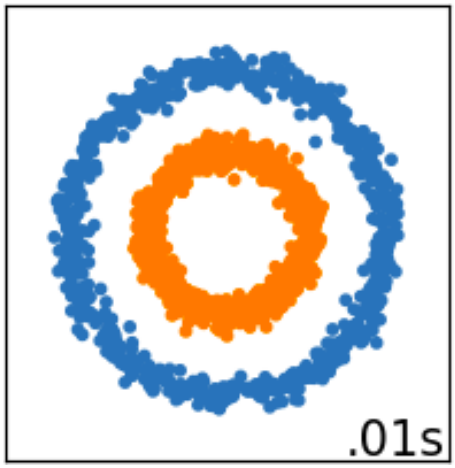


Figure 2.5: Example of DBSCAN algorithm- Reproduce from Scikit-learn [20]

2.2.5 Evaluation

In the evaluation phase is the one that we compare the modelling results with our objectives defined at the beginning(business understanding). If the results aren't satisfying we should go back to the first phase to understand what went wrong or to determine a new set of objectives.

2.2.6 Deployment

The deployment phase is when we put the model that we have in the production, usually this moment arrives when the result of the evaluation phase was good or at least satisfying.

2.3 Automated Fare Collection (AFC) Systems

Automated Fare Collection systems are quite usual nowadays in almost every public transport around the world, and is also implemented in the public transport of the metropolitan area of Porto.

There are two different types of AFC system, entry-only and entry-exit systems. In the entry-only system it is only necessary to tap the ticket in the reader on the entry of public transport, while in the entry-exit system is necessary to tap on the entry, but also on the exit. This has a huge impact on the dataset that we can derive from these systems, because entry-only systems provide the origin of passengers journeys, whereas entry-exit system provides the origin and destination of the passengers. The AFC system in the study is Andante, and it is an entry-only system. Several studies use data retrieved from these types of systems, as shown in Table 1. Most of the studies use data from Entry-only systems, probably because most of the systems are Entry-Only, typically only in trains and metros we see the utilization of the Entry-Exit System.

These studies have different approaches to retrieve valuable information from the data. The most common way of valuable information, normally from these studies[11, 3, 1, 15, 9, 13] is OD-Matrix.

Papers	AFC System - Entry-Only	AFC System - Entry-Exit
[11]	X	-
[3]	X	-
[1]	X	-
[4]	X	X
[15]	X	-
[10]	X	X
[3]	X	-
[9]	-	X
[13]	X	X

Table 2.1: Studies about the different AFC Systems

2.4 OD-Matrix

The deduction of destinations can result in an OD-Matrix(origin-destination matrix). The OD-Matrix can be quite useful to understand and discover mobility patterns of STCP users that is the aim of this study. OD-Matrix, in this type of studies, is a matrix with the origins and destinations of users that is deducted from the public transport data. Normally, the part that is deducted is the destinations, because, as we can see in the Table 2.1, most of the studies are entry-only systems consequently there is a need for deducting the destinations. In this section, we are going to talk about some solutions that already were used to help the deduction of an OD-Matrix, and studies that done this matrix.

Barry et al. [1] and Barry, Freimer, and Slavin [3] are studies that occurred in New York City, where they deducted the destinations, but also, in some cases, the origins[3]. Most of these studies had to make some assumptions on the processes of extracting valuable information from their data. Those assumptions are:

- The exit station of a user bus/train trip is most likely the entry station of the next bus/train trip.
- The entry station on the first trip of the day is most likely the exit station of the last trip of the previous day.

The majority of the studies[1, 3, 11, 9] use these assumptions because without them would be very hard to deduct the exit points of the users. These assumptions had 90% of valid destinations in a survey done by New York City Transit(NYCT) to validate the destinations of their methodology [1].

Barry et al. [1] made an OD-Matrix for their subway system, the MetroCard. They used those assumptions made before deducting the OD-Matrix and confirmed that their methodology had 90% of valid destinations, by this, they mean that 90% of their data respect those assumptions making it possible to deduct destinations. They reach that result by dividing their users into two groups and then using a sample of each one. The first group is composed of users that in one day only had two trips, and the second group is composed of users that had more than two trips. In

each group, they created different possible use cases that could occur in their data. In group 1 they had 4 cases represented in 2.6 and group 2 they had 3 cases represented in 2.7, both with the respective results.

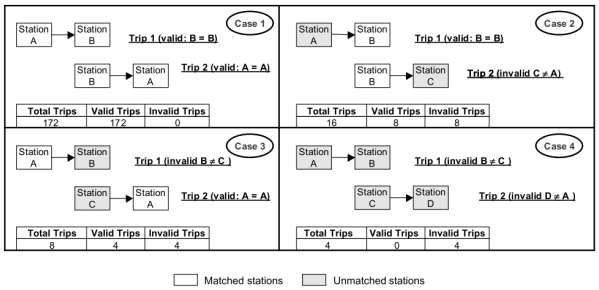


Figure 2.6: Validation of group 1 data - Reproduce from Barry et al. [1]

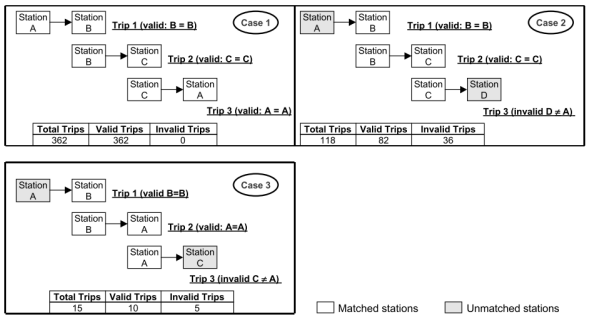


Figure 2.7: Validation of group 2 data - Reproduce from Barry et al. [1]

After, their OD-Matrix deductions were compared to the counts made in certain stations and the results were quite good to be fair. Their predictions for *60th Street* were 123,000 and the counts had 128,000, for *Queens* were 103,000 and the counts had 105,000 and for *Brooklyn* were 152,000 and the counts 150,000.

This is an example of how much these assumptions can be good for this type of studies, mainly because they represent well what happens in reality.

2.4.1 Euclidean Distance

The Euclidean distance is used to define a distance, like a perimeter, that a user can go through from the station that they come off to other different station that they need to go to pick another public transport. This helps to deduct OD-Matrix by defining a maximum distance that a user can go through from the "exit" to another "entry" of public transport. For example, a user comes off in the bus station and can go to a different bus station around that place, or even go a metro station, this is the reason some studies use Euclidean distance [11, 9]. In these examples, they use a different set of values to the interchange distance, Nunes, Dias, and Falcão E Cunha [11] utilized 400m, 640m and 1000m, where at Alsger et al. [9] 400m, 800m, 1000m and 1100m were utilized. In Nunes, Dias, and Falcão E Cunha [11] the 400m example, he had 10.5% of the

transaction records failing the interchangeable distance, in the 640m he had 7.7% of the records failing and in the 1000m he had 5.8%. These results were associated with their influences on the possible destinations inferred, when Nunes, Dias, and Falcão E Cunha [11] utilized 400m he could infer 60% of the transactions, and when compared with the 640m that had 62.4% and the 1000m that had 64.6%, he decided that the 640m was the best choice because the 1000m example could have false-positives, accepting an incorrect candidate destination of the transaction, and 400m could reject some true positives, rejecting the correct destination of the transaction. Looking for the results that he had the 640m was a wise choice for interchangeable distance considering that 1000m example cut the failed transactions almost by half when compared to 400m, that is a big cut of the failed transactions that could mean that he was accepting transactions that are false-positives as was mentioned before, and the 400m had a percentage of failed transactions too high.

Alsger et al. [9] as mentioned before is an Entry-Exit system, and they try to deduct the destinations of each trip utilizing trip-chaining with some transfer rule. Those rules are that a transfer can't exceed x time and y distance. If it exceeds that last stop is considered a destination and not a transfer. Alsger et al. [9] used some values different than Nunes, Dias, and Falcão E Cunha [11]. This time the values used for walking distance were 400m, 800m, 1000m and 1100m in conjunction with the transfer time of 30min, 60min and 90min. The estimations of destinations in this study were compared with OD-Matrix derived from real data, that utilized transfers times to define if the trip destination was indeed a transfer or a destination. The results of the estimations had, with all the different type of values for walking distance and time, more OD-Trips than the actual OD-Matrix from the real data. Probably their estimations were deducting different trips that in reality was one, and this can come from the fact that the estimations have a distance limit, and the OD-Matrix from the real data don't, only time. Their results also showed that 80% of their transfer trips have an average transfer walking distance of 400m at all the different transfer times[9].

These studies [11, 9, 1, 3] all use OD-Matrix on their approaches to the data from public transport. All of them tries to deduct destinations to create an OD-Matrix, even [9] that has the real exits tries to deduct the destinations utilizing their real destinations has validation to train the model that deducts destinations. This approach of making an OD-Matrix probably to better understand the mobility patterns will be dependent on the availability of data for the realization of this approach. This specific type of data consists of having a certain SmartCard(unique ID) going back and forward in the same places and from that we can deduct the OD-Matrix. In the context of this study, the data is only from non-frequent passengers(Average of 30% of all the data). Most of these non-frequent passengers are sporadic users and their data probably will not respect the assumptions that were talked at the beginning of this section(Section III), because most of them are single trips or as said before, they are used sporadically what makes it hard to deduct something from that.

2.5 Clustering based Approaches

In this section, we will talk about some approaches where clustering was applied and their correlation with our objectives for this study.

El Mahrsi et al. [12] made two different clustering approaches to the data. First, they made an approach where they applied the technique on the stations, and the other approach they applied on the passengers based on their temporal behaviour. The first approach will aggregate the stations that have similar usages profiles [12], so each cluster will represent a certain profile of station.

El Mahrsi et al. [12] on their approach to stations they utilized an adaptation of Bicycle Sharing System station clustering model reported in Etienne and Latifa [7]. The approach that these studies used is a Poisson mixture model. In these studies, both of them used this model, and both of them used EM (Expectation Maximization) algorithm for the tuning of the parameters of Poisson mixture model, more specifically, the search for the best number of clusters that they could use in each case. They reach that desired number of clusters by finding the knee/elbow of the EM algorithm, and because sometimes that knee/elbow change is not simple to determine, they utilized a linear model fitted to the linear part of the curve to find point[12]. The details and advice for the implementations of this linear model are in [5]. Fig. 2.8 shows the knee/elbow point, and the numbers of clusters that they worked on in[12].

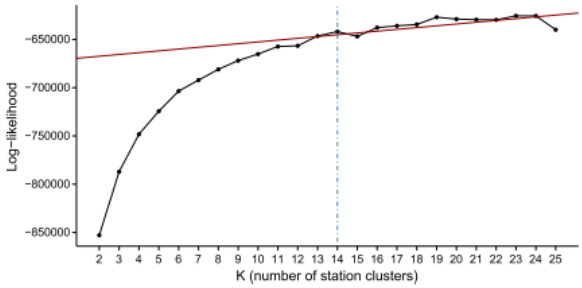


Figure 2.8: Evolution of the log-likelihood as a function of the number of station clusters K = 2,. .., 25 (EM algorithm). (Red Line) The linear model fitted to the linear part of the curve. (Blue Vertical Line) Suitable number of clusters K = 14. - Reproduce from [12].

El Mahrsi et al. [12] had another interesting approach. This time their objective was to discover the passengers that presented similar temporal boarding times. They made interesting use of the mixture unigrams model. Usually, this model is used to cluster documents in the context of information retrieval but they used it to cluster the passengers boarding times. In their study, they explain what they did to the data to make more suitable the use of this model [12]. As before, in this approach, they used the EM algorithm for the tuning of the cluster's parameter.

The results of each approach were interesting. The station approach results showed that there are different patterns related to the place where the stations are located. Those patterns are for example "housing" stations, stations that are close to houses, or work stations, stations that are close to workplaces. These are the main patterns that are visible on their results, and they can be useful to know if there are people that use tickets to go to work or people that use tickets to visit

points of interest. Another interesting information that could be retrieved from this approach is to know which are the peak hours of use in certain stations for non-frequent users.

The passenger's approach is interesting, but for us, their conclusions[12] aren't that useful because they use the passengers boarding times in conjunction with the fare types. In our study we just have the fare type that says ticket for the non-frequent users(that are the type of passengers that we are looking for), so we could use their approach not in conjunction with the fare types, but to discover the type of users by looking to their boarding times and look for patterns. Consequently, if a user has a diffuse pattern of utilization probably is a tourist, or if they have a strict use, one time at 7 am and other at 5/6 pm probably it's a local user that utilizes tickets to work. This type of information could be useful to understand the type of person that is using tickets instead of a subscription. An additional approach that we could apply from these approaches is to associate both of them to retrieve more information. Hereby the association of those approaches results could lead us to the type of users that uses each type of station. For example, a tourist could have a more diffuse pattern when looking to the times that he uses the public transport, but the stations that he uses could be more likely to be near to points of interest. We could retrieve this type of deduction by associating those results but to be sure just after doing it.

Zhao et al. [13] analysed their data in different perspectives to reach their goal that was to investigate passengers travel patterns and discover regularity and anomalies in those patterns. Their perspectives of how to analyse and search for patterns in data are divided by temporal travel pattern, spatial travel pattern and spatio-temporal patterns. The extraction of each pattern had their its methodology to organize the data. Their patterns were utilized to "feed" the clustering algorithms.

The temporal travel pattern consists of looking for the periods that a passenger performs his trips. The representation of temporal travel pattern for each passenger is made by an ordered set $\tau = \{T_1, T_2, .., T_n\}$, and each Ti have associated a $Ti.t$ that is a time slot and $Ti.r$ that represents the proportion of the number of active days on that time slot. The time slots in τ have 3 hours length and do not overlap, for example, $T_1.t$ and $T_2.t$ for a certain passenger are 7:00 9:59 and 18:00 20:59. Also, these slots are ordered by the active time of the passenger on that slot. In the procedure of making this representation the overlapping of the times slots is done to cover for example the case 9:30 11:00. There are more details in [13]. The spatial travel pattern here is represented by $S = \{S_1, S_2, ...\}$ basically an OD Matrix, with the number of trips and the proportion of the number of active days for the OD pair as well. These information are in each Si, $Si.o$ and $Si.d$ are the OD pair, $Si.n$ is the number of trips and $Si.r$ is the proportion [13].

2.5.1 Spatio-Temporal Patterns

The spatio-temporal patterns of a passenger are represented by $\mathscr{R} = \{R_{i,j}\}$. The spatio-temporal patterns item $\{R_{i,j}\}$ have information about the relationship between the temporal pattern T_i and the spatial travel pattern S_j. Those information details are in [13].

As we've talked before, the search of these patterns are very similar to the possible idea that we retrieve from the [12] approaches. The association of hour of utilization with space of utilization.

The algorithm utilized to cluster passengers was K-means. The search for the optimal number of clusters was done with the help of the Silhouette coefficient. The silhouette coefficient value is used to measure of how similar a passenger is to the cluster (cohesion) compared to others (separation). The silhouette coefficient is in [-1, 1], where a large value indicates that the passenger is well matched to the cluster and poorly matched to others[13].

The results for both cases were quite similar, and that was confirmed by their correlation afterward. For each clustering results there are two clusters that represent some significant pattern in the data. Those two clusters are persons that have one or two OD pairs respectively in the case of spatial pattern. In the temporal case, those two clusters are persons that have one and two time slots respectively. This means that for the temporal case there are passengers that use metro regularly one time and others two times. The same for the spatial case. Passengers that have only one OD pair, and others with two OD pairs.

The results talked above were from the metro data. Afterward, they added bus data to the metro data and rerun the clustering algorithms. Their findings were quite interesting because in both cases the cluster that had one OD pair and one time slot almost disappeared and went to the cluster of two OD pairs and two time slots respectively. This means that they are people who use the metro at one point of the day and the other they use buses. They find that this was mainly because bus cost less than the metro, and was easier to find a source to destination with a bus than metro [13].

The way that the authors [13] analyzed the metro data and gather some valuable information is interesting for us, because we have pretty much the same type of data available, the only problem will be the OD pairs because our dataset has only the non-frequent passengers and the determination of the OD pairs could be unfeasible. Instead of that, we could just look for the station of boarding and see the results of that. Another perspective that they had was to aggregate the bus data something that we have too.

Another approach could be doing these approaches, temporal and spacial for each type of transport, in this case, bus and metro. Afterwards, if there is some correlation, or not, in the patterns, or associate the results of each type of transport to find things like [13]. That find that people used the bus because it was cheaper and easier to go for a specific place as we talked above.

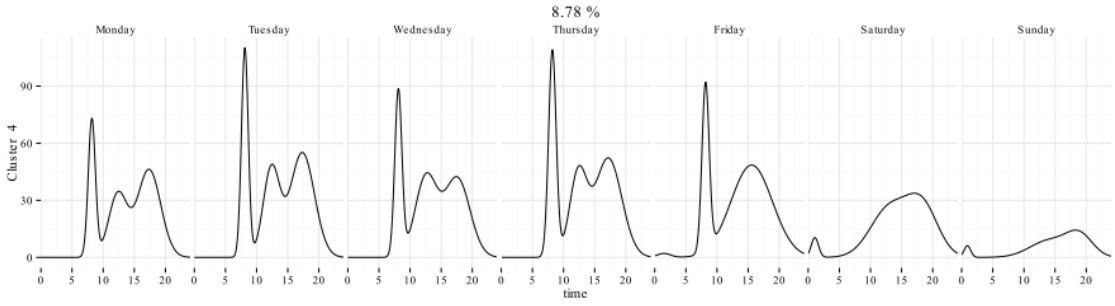
In Zhao et al. [13] there are some analyses and distributions that were interesting to look at, and that helps to better represent what does the data mean. Those analyses and distributions are mostly presented in figures on [13].

Briand et al. [8] is another study that does the clustering of the temporal passenger profiles. In their work, they do this approach of clustering temporal passenger profiles from the real ticketing data collected during April 2014. Their data is from the STAR (Service de Transport en Comun de l'Agglomération Rannaise) public transport network of Rennes Métropole(France)[8].

Their clustering approach to cluster the temporal passenger profiles is a two-level gaussian mixture model. In their model, their first level consists of clustering passengers based on their temporal activity, and in the second level clustering by the ticketing logs, mostly based on the timestamp information[8]. In other words, their first clustering level looks only for the times of

activity of passengers and clusters by that, and in the second level, they combine those with each passenger trip, to make a more detailed cluster that represents how the trips made are distributed over time[8]. They explain their methodology and implementation of the model to achieve this, in Section IV [8]. In this section, we can also see that they use CEM and EM algorithm for model calibration. These methods are the most commonly used methods for mixture model estimation[8].

One of the key aspects that I liked from this study is how their results are so clean and easy to understand. Figure 2.9 is an example of a clustering result. As we can see their graphs are so easy to understand and consequently easy to retrieve good information from it. For example, in fig.2.9, we can see that most of the weekdays have three peaks of activity, but also there is another interesting peak at night time that occurs only on Friday and Saturday. This result is interesting because it shows how much this model can find new activity patterns on it.



(a) Temporal profil

Figure 2.9: Example of one Temporal profile result from Briand et al. [8] results - Reproduce from [8]

Papers	Temporal Profile	Spatial Profile	Spatial-Temporal	Model	Model Calibration	Statistics
[12]	X	X	-	Poisson mixture model (Spatial) and Mixture unigrams model(Temporal)	EM algorithm	X
[13]	X	X	X	K-means model	Silhouette Coefficient	X
[8]	X	-	-	Two-level Gaussian mixture model	CEM and EM algorithm	X
[14]	X	X	X	*Louvain* algorithm[2]	-	X

Table 2.2: Characteristics of each study about the different AFC Systems

Liu, Gao, and Xin [14] is a study that measures the diversity and dynamics of urban mobility patterns of the municipality of Chongqing, from China. This paper uses the smart card data collected by the subway system of Chongqing. The portion of data collected to analyze is from one

week(from 04/08/2014 to 10/08/2014).

In this paper, they study the diversity and dynamics of urban mobility patterns from three different aspects. The first aspect is from individual travel behavior. In this aspect, they plot the number of trips made by all the passengers in each hour of the day that week. One of the obvious patterns on that plot is the difference in usage between the weekdays and weekends. Afterwards, they utilized the Pearson correlation matrix[6] on the data to measure the correlation between the different days. In section 3.1 of the paper, they specify the calculation method of the Pearson correlation coefficient used[14].

The second aspect is from geographical locations. This time they plot the temporal usage of each station. The plots for this aspect are composed of the trips done in each hour at a station. Afterwards, they used Pearson's correlation again. In this aspect, they correlated the usage of each station by days. This means that they have a correlation matrix of the usage of stations for each weekday. They find with this that stations with close ids, that are geographically close, have similarities on usage.

The third aspect looks for the diversity and dynamics of the crowd flow network. In this aspect they use, a famous community detection algorithm, *Louvain algorithm*[2] to cluster the nodes of crowd flow network[14]. They analyze the community patterns on a temporal scale to better understand and measure the diversity and dynamics of the crowd flow network. So they plot the changes that occur in the community clusters in each hour of the day that they picked.

All these aspects that they approach have important results in them. In the first aspect we can figure out the overall usage of the transport, in the second we can see what are the stations that are more used and which ones have similar patterns and the last aspect of the crowd flow is useful to understand which stations are having a similar usage on a specific hour, and how that varies over the day. The implementation of these aspects on our project is feasible, at least the first and the second, and from that, we can figure out already some patterns of utilization of the passengers. The last aspect is harder to be sure if it is feasible due to our dataset not having the destinations and the deduction of those being depended on some conditions as we speak earlier in Section OD-Matrix.

So far most of these studies talk about clustering the data from an AFC system. They have different approaches that they use on their data. Most of these studies try to profile passengers by temporal, spatial or spatial-temporal activity. Table 2.2 aggregates the purpose of each clustering algorithm used, in each study.

Chapter 3

Problem Characterization

3.1 Domain

The metropolitan area of Porto(MAP) is composed of 17 municipalities, whereas one of these municipalities is the city of Porto that is the second biggest urban area of Portugal. In the MAP there is an entity called *Transportes Intermodais do Porto* (TIP) that was constituted by Metro do Porto, SA, pela Sociedade de Transportes Colectivos do Porto (STCP), SA e CP- Comboios de Portugal at 20 of December of 2002, to promote the implementation of inter-modality in public transport [16]. This entity created the ANDANTE system. The ANDANTE is an Automated Fare Collection (AFC) system that handles all the ticketing and fares of the transport operators operating in MAP and that is part of the Sistema Intermodal Andante(SIA) that are in it. Those transport operators are:

- Metro do Porto
- STCP
- Comboios de Portugal (CP) - 4 lines
- Maré de Matosinhos
- Espírito Santo - Autocarros de Gaia
- Maia Transportes
- Valpi Bus
- ETG - Empresa de Transportes de Gondomar
- MGC - Transportes
- Auto-viação Pacense
- Auto-viação Landim

- Transdev
- Arriva
- RodoNorte
- Caima
- Acosta Reis
- Seluve
- UTC

These transport operators are composed of different types of transport. They are buses, light rail and suburban trains. These transport operators cover all the zones in the MAP to reduce car traffic in the metropolitan area of Porto.

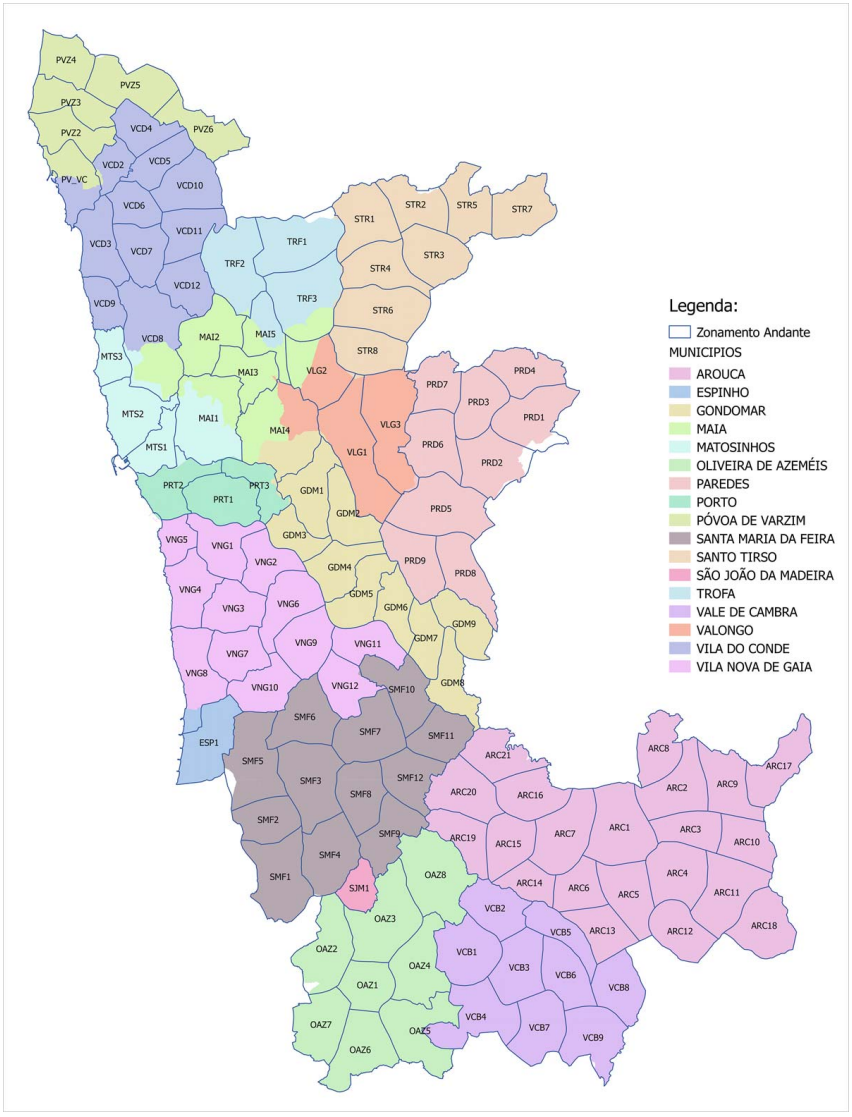
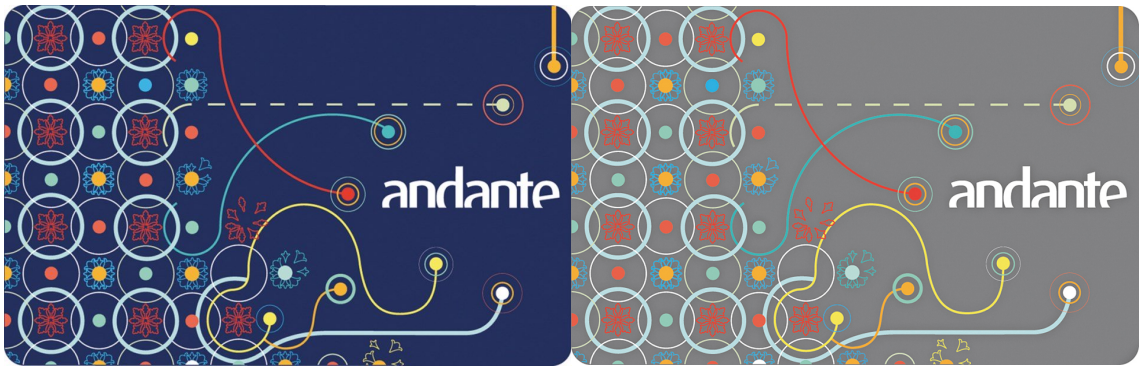


Figure 3.1: Zones of Metropolitan Area of Porto



(a) Occasional Ticket (b) Monthly Pass

Andante has two types of tickets. The occasional ticket(Fig.3.2a) and the monthly pass(Fig.3.2b). Since MAP is organized by different travel zones, as figure 3.1 shows, Andante is a distance-based system. This means that for occasional tickets $Z_i, i \in 2, ..., 12$ represent the number of zones that a user is allowed to travel in a single journey and for monthly passes they represent the number of zones that an user can choose to use in Andante. Andante is also a time-based system. The users can make unlimited transfers during a period an will only use one charge. The next tables represent the prices for occasional tickets and monthly passes.

Zone Type	Price (Euros)	Time Allowed	Andante 24(Ticket for 24 hours)
Z2	1.20	1h	4.15
Z3	1.60	1h	5.50
Z4	2.00	1h15min	6.90
Z5	2.40	1h30min	8.30
Z6	2.80	1h45min	9.65
Z7	3.20	2h	11.05
Z8	3.60	2h15min	12.40
Z9	4.00	2h30min	13.80
Z10	4.40	2h45min	15.20
Z11	4.80	3h	16.60
Z12	5.20	3h15min	18.00

Table 3.1: Occasional Ticket Fares

Zone Type	Normal	Social	Social+	4_18Sub23
Z2	31.15	23.85	15.60	12.45
Z3	38.40	28.80	19.20	15.35
Z4	48.65	36.50	24.35	19.45
Z5	58.85	44.15	29.45	23.55
Z6	68.60	51.45	34.30	27.45
Z7	78.35	58.75	39.20	31.35
Z8	88.10	66.10	44.05	35.25
Z9	97.85	73.40	48.95	39.15
Z10	107.60	80.70	53.80	43.05
Z11	117.35	88.00	58.70	46.95
Z12	127.10	95.35	63.35	50.85

Table 3.2: Monthly Pass Fares

There are also the occasional ticket "Andante Tour" that is the ticket that has unlimited travel charges and allows to use all the MAP. They are only available in two options. They are:

- Andante Tour 1 - That is valid up to 24 hours after the first validation
- Andante Tour 3 - That is valid up to 72 hours after the first validation

Also, in April 2019, a tariff reduction support program (PART) was imposed. This had the purpose of attracting more and new customers to public transport. In this program, transport authorities were supported with funding to reduce the tariffs and create better conditions to access the public transport network. So, the new tariffs applied are presented in the next table.

Zone Type	Normal	SOCIAL*/4_18/SUB23	SOCIAL+(A)	4_18(A)/Sub23(A)
Up to 3 zones	30.00	22.50	15.00	12.45
More than 3 zones	40.00	30.00	20.00	16.00

Table 3.3: Monthly Pass Fares

3.2 Problem Formalization

The public transport network as a huge impact on the satisfaction of their customers(passengers). This satisfaction normally correlates to the way that the transport network corresponds to their customer's needs. This satisfaction could be easily guaranteed by having transport passing in intervals of 5 min, on every single route of the network. That is expensive and unfeasible. Also, having transport passing in intervals of 1 hour, in every single route would make a lot of passengers not pleased with the transport service. However, it is cheaper and feasible.

The problem is that passengers flow in the network isn't always the same. This means that every route in the network could have different schedules to better suit the passenger's needs. So,

if there are many passengers of a specific location and time that utilize the transport this could mean that a specific route could need more transport passing. In the other hand, if a certain location and time of the day don't have the affluence to the transport this could mean that a cut on the transport passing on that route and time makes no difference to the passenger's needs.

Several factors can impact the passengers flow in the network. They are:

- **Day of the week** - Depending on the weekday the passenger's flow in the network can change. For example, when comparing a Monday to a Sunday the differences are obvious, the affluence to the transport on Sunday are fewer than on Monday;
- **Month/Seasonality** - The seasons of the year can have a direct impact on the affluence to the transport. Comparing Summer to Winter. Normally, in Summer there are more tourists in town and that can have a direct impact on the non-frequent passengers using transport;
- **Hour of the Day** - The hours of the day can impact the passengers flow to the transport network. Normally, in frequent passengers, there are two spikes on the affluence to the transport that are the morning hours before work, and the evening hours after work. This happens to frequent passengers, and what happens to non-frequent passengers?;
- **Location** - The location has an impact on the flow. There are spots on the network that have naturally more affluence than others because they are more central or are closer to shopping or points of interest.

3.3 Proposed Solution

The solution that we came up to address this problem and understand better the way that those factors influence the transport network is by searching three types of patterns. They are temporal patterns, Spatial Patterns and Spatio−Temporal Patterns.

The temporal Patterns will be our focus on the behaviour of passengers in the transport network related to time. Time related patterns to profile the passengers by the time of transactions(time of transaction into the transport). These profiles can be done by day, week, month, morning hours, evening hours or late hours. These profiles of time of transactions help to retrieve knowledge from some of the factors represented above, such as day of the week, seasonality and hours of the day. These time of transactions profiles can show the impact that those factors have. Representing the profiles of each day of a week and look for similarities or differences is an example of what we can do to show the impact of a weekday.

The Spatial Patterns will be our centre of attention on the behaviour of passengers in the transport network related to location. Spatial patterns will reveal the locations that have more affluence by the passengers, in the transport network. Those locations can be related to a day, week, month or season. This information about the spatial patterns of the transport network extracts knowledge related to all the factors showed before. This means that we can characterize how each factor impacts the spatial characterization of the network. Representing the locations with more affluence

in each day of the week or the locations with more affluence in a random week of each month or the locations with more affluence in the workdays compared to the weekend. These are some examples of how to extract spatial knowledge from the data.

The Spatio−Temporal Patterns are going to be our focus on the behaviour of passengers in the transport network associated with time and location. These Spatio−Temporal patterns will, most certainly, reveal the locations and their respective time of more affluence and represent the pattern of utilization in the transport network. These patterns can be comparable by day, week, month or season. These patterns also can extract knowledge of how each factor impacts the network in terms of space and time. Representing the stations that have more affluence in a specific hour, or the stations that have similar patterns of utilization. These are some examples of what these patterns can be used to extract spatio-temporal knowledge from the data.

Chapter 4

Exploring the Andante AFC data for unfrequent users

The main methodology used in this study was CRISP−DM. We used it throughout all study in all types of patterns. First, we tried to understand in terms of business what could be useful information in all types of patterns, for example, understand what could be valuable information in Temporal Patterns, like creating profiles of utilization. Then we would make an understanding of the data and prepare that data to reach the objective defined before, in this case, we would look for data that has temporal information and aggregate only that type of data. Afterwards, we would use that data to create a model and discuss that model results, for instance, understand if the temporal profiles make sense and if it is a pattern.

This was a back and forward methodology. If we found something new in terms of business understanding that could have valuable information, we will make all the process of the methodology to reach the results.

Figure 4.1: CRISP-DM Methodology process

4.1 Introduction

This chapter is divided into five parts. Data description, data pre-processing, and presentation of the three main patterns that were studied Temporal, Spatial and Spatio-Temporal patterns. In each section it will be represented all the data mining techniques that were used to extract knowledge from each specific type of patterns.

4.2 Data Description

In this section, data used in this study is presented along with the dataset format and size. Also, we will present some findings relate to occasional users, in this case, the new users per day.

The dataset in the study is from Andante AFC system and it relates to the 2013 year. This dataset is composed of 12 text files. Each one of these files is composed of the validations done in each month of the 2013 year. Each file has an average size of 1 Gigabyte. This corresponds to an average of 10 million transactions per file. These transactions are all the transactions done in the network, on each month. The following image is a portion of January dataset to illustrate the data.

```
op,cod_cartao,Paragem,Linha,Sentido,Variante,Veiculo,zonamento,dhiniviagem,datahoravalidacao,Tip_Valid,Grupo_Cod,Grupo,Perfil_Cod,Perfil,zonas,
STCP,20007188034,4207,907 ,1,1,3032,2,2013-01-30 19:37:53,2013-01-30 20:19:19,3,2,Assinatura,100,Normal,C1-2 S8-9,S9
STCP,20031228054,2211,907 ,2,1,3032,2,2013-01-30 20:36:49,2013-01-30 20:36:24,3,2,Assinatura,100,Normal,C1 S8-9,S9
STCP,20031312314,2085,907 ,2,1,3032,2,2013-01-30 20:36:49,2013-01-30 20:37:30,3,2,Assinatura,107,Estudante,C1 S8-9,S9
STCP,20031259162,2085,907 ,2,1,3032,2,2013-01-30 20:36:49,2013-01-30 20:37:32,3,2,Assinatura,30648,Social+,C1 S8-9,S9
STCP,20032224996,2085,907 ,2,1,3032,2,2013-01-30 20:36:49,2013-01-30 20:37:14,3,2,Assinatura,100,Normal,C1 S8-9,S9
STCP,20032171602,2085,907 ,2,1,3032,2,2013-01-30 20:36:49,2013-01-30 20:37:16,3,2,Assinatura,100,Normal,C1 S8-9,S9
STCP,20007315826,2184,907 ,2,1,3032,2,2013-01-30 20:36:49,2013-01-30 20:38:27,3,2,Assinatura,100,Normal,C1 S8-9,S9
STCP,20032152919,2184,907 ,2,1,3032,2,2013-01-30 20:36:49,2013-01-30 20:38:31,3,2,Assinatura,30698,Social+ R,C1-2 S8-9,S9
STCP,20030239067,2278,907 ,2,1,3032,2,2013-01-30 20:36:49,2013-01-30 20:39:30,3,2,Assinatura,30648,Social+,S2 S8-9,S9
STCP,20007387937,2278,907 ,2,1,3032,2,2013-01-30 20:36:49,2013-01-30 20:39:14,3,2,Assinatura,100,Normal,S8-9,S9
STCP,20007340254,2278,907 ,2,1,3032,2,2013-01-30 20:36:49,2013-01-30 20:39:36,3,2,Assinatura,100,Normal,S8-9,S9
STCP,20030171068,2046,902 ,2,1,3215,2,2013-01-31 20:38:16,2013-01-31 20:47:09,3,2,Assinatura,100,Normal,S1-2 S8,S1
STCP,20032158114,3599,902 ,1,1,3215,2,2013-01-31 21:23:06,2013-01-31 21:23:55,3,2,Assinatura,107,Estudante,C1-3 S8,C1
STCP,20007391470,3599,902 ,1,1,3215,2,2013-01-31 21:23:06,2013-01-31 21:23:50,3,2,Assinatura,100,Normal,C1 C6 C9 S8-9,C1
STCP,30139203085,3599,902 ,1,1,3215,2,2013-01-31 21:23:06,2013-01-31 21:23:57,3,1,Título de Viagem,100,Normal,Z3,C1
STCP,20032206734,3599,902 ,1,1,3215,2,2013-01-31 21:23:06,2013-01-31 21:23:51,3,2,Assinatura,107,Estudante,C1 C6 S1 S8,C1
STCP,20032163256,3599,902 ,1,1,3215,2,2013-01-31 21:23:06,2013-01-31 21:23:54,3,2,Assinatura,107,Estudante,C1 S8,C1
STCP,20031333759,3599,902 ,1,1,3215,2,2013-01-31 21:23:06,2013-01-31 21:24:00,3,2,Assinatura,100,Normal,C1-2 S8,C1
```

Figure 4.2: Portion of Raw data from January 2013.

The 4.1 table shows the size and the number of transaction in each month file.

Month	Size	Transactions
January	1.2GB	11853948
February	1.0GB	10478998
March	1.1GB	11084125
April	1.2GB	12199772
May	1.3GB	13043034
June	1.0GB	10620507
July	1.1GB	11246006
August	0.9GB	8971762
September	1.1GB	11178624
October	1.3GB	13113395
November	1.1GB	11556280
December	1.0GB	10345151

Table 4.1: Data Files Size and Transactions

Every single transaction is composed of several columns of information. Those columns and their respective information are:

Column	Information	Type of data
op	Name of the public transport operator	string
cod_cartao	Numeric ID of Andante Smart Cards	numeric
Paragem	ID of stops	numeric
Linha	ID of public transport lines	string
Sentido	Direction	string
Variante	Type of bus line (short/long)	string
Veiculo	Numeric ID of vehicles	string
zonamento	N/A	numeric
dhiniviagem	Date and time in which a trip started	AAAA-MM-DD hh:mm:ss
datahoravalidacao	Date and time of an Andante Smart Card transaction	AAAA-MM-DD hh:mm:ss
Tip_Valid	N/A	string
Grupo_Cod	Numeric ID of ticket categories	numeric
Grupo	Text description of ticket categories	string
Perfil_Cod	Numeric ID of user profiles	numeric
Perfil	Text description of user profiles	string
zonas	Zones for which the Andante Card used is valid	string
Zona	Zone where validation occurred	string

Table 4.2: Columns Information

4.3 Pre-Processing

This section discusses all the techniques needed to correct the raw dataset. Afterwards, it will be also shown in each pattern the pre-processing techniques and changes to the data that were required.

4.3.1 Data Reduction and Cleaning

Since this study is only about the non-frequent passengers, our first move was to make a data reduction. This data reduction consisted of removing the data from frequent passengers. To achieve that, a verification of the value in the column "Grupo" of each dataset entry was done. If this column had any of the following values: "Assinatura", "Assinatura STCP" or "Assinatura Fim de Semana STCP", they were removed because these are entries of frequent passengers. This means that we only keep with data related to non-frequent passengers. Afterwards, data cleaning was needed because the dataset had different stations with the same id. This happens because there were bus stations and metro stations with the same id. So, for the Andante AFC system, it was given a unique id for these different transport stations. This information about the new ids related to the old ids is in the station's dataset(Figure 4.4). As a consequence, a cleaning of those wrong

ids were needed, and to accomplish it, it was done a comparison, in each entry of the transactions dataset with each entry of the stations dataset. Then, when the columns "op" and "paragem" from the transactions dataset had the same values, respectively, of the columns "op" and "oldCode" of the station dataset. We changed the value of the column "Paragem" in the entry of transaction dataset with the value of the column "newCode" of the entry in the station dataset that had the same values in those columns. Furthermore, we also added in each entry of the dataset the latitude and longitude of the station as two new columns "lat" and "long". The following two images show a portion of the stations dataset(Figure4.3)and a portion of transactions dataset in the new format(Figure4.4).

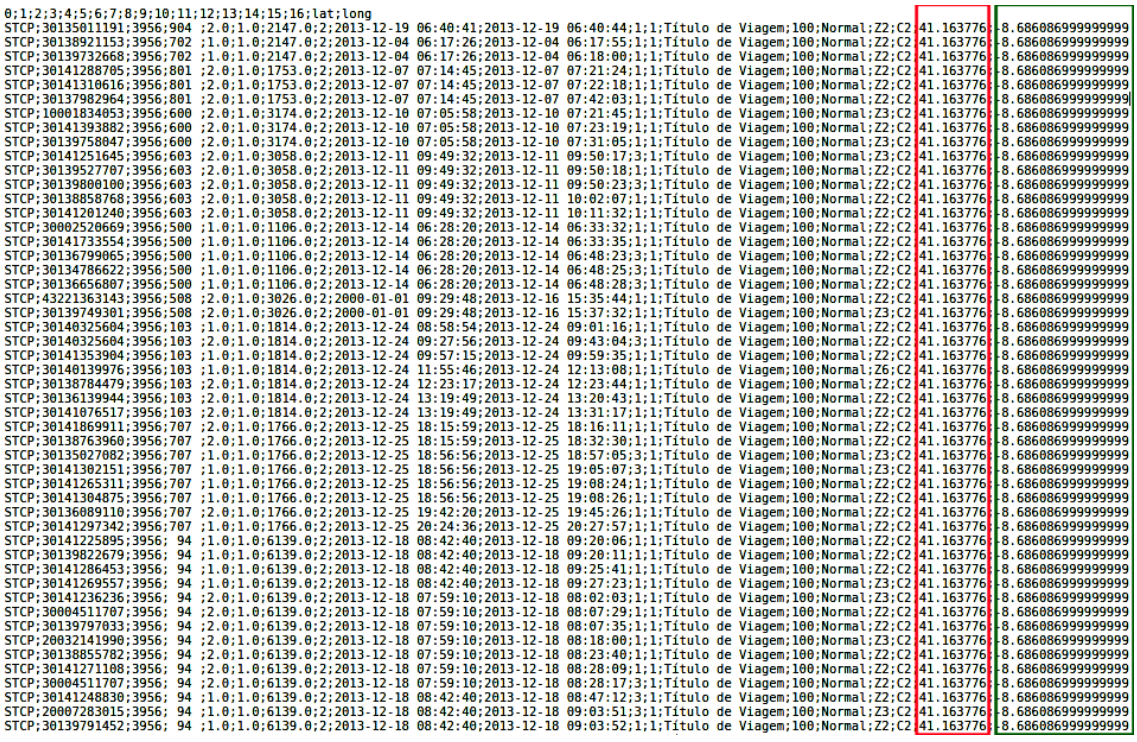


Figure 4.3: Portion of the Transactions dataset with the two new columns "lat"(red) and "long"(green)

```
newCode;op;oldCode;cod_o;d_plaragem;c_stop;zona;lat;lon
3956;STCP;0;0;Homem do Leme;HMLM3;C2;41.163776;-8.686087
3957;STCP;0;0;CASA DE RAMALDE;CRM3;C2;41.1690744;-8.6463037
3958;STCP;0;0;S. JOÃO BOSCO;SJBC3;C2;41.166899;-8.6479415
3959;STCP;0;0;PR. DE LIÉGE;LIEG1;C2;41.1552084;-8.6783877
3960;STCP;0;0;MOLHE;MLH5;C2;41.160063;-8.6822452
3961;STCP;0;0;PR. DE LIÉGE;LIEG2;C2;41.15498;-8.6781985
3962;STCP;0;0;DIU;DIU1;C2;41.153655;-8.6764405
3963;STCP;0;0;PASSEIO ALEGRE;PASS1;C2;41.148996;-8.6732342
3964;STCP;0;0;DIU/TEATRO;DTR;C2;41.153275;-8.6770352
3965;STCP;0;0;PÊRO VAZ CAMINHA ;PVC1;C2;41.157482;-8.6675942
3966;STCP;0;0;PÊRO VAZ CAMINHA ;PVC2;C2;41.157314;-8.6675352
3967;STCP;0;0;FOZ;RFAR1;C2;41.152784;-8.6778282
3968;STCP;0;0;PÊRO VAZ CAMINHA ;2PVC2;C2;41.157302;-8.6678288
3969;STCP;0;0;GUERRA JUNQUEIRO;GRJ;C2;41.155927;-8.6380988
3970;STCP;0;0;PASSEIO ALEGRE;PASS3;C2;41.149141;-8.6727988
3971;STCP;0;0;PASSEIO ALEGRE;PASS4;C2;41.149036;-8.6732713
3972;STCP;0;0;SETE BICAS;SETB;C2;41.181913;-8.6536468
3973;STCP;0;0;MOSTEIRO LEÇA DO BALIO;MTLB1;C5;41.210544;-8.6253663
3974;STCP;0;0;RHMAIS-LIONESA;LION1;C5;41.214343;-8.6244822
3975;STCP;0;0;MOSTEIRO LEÇA DO BALIO;MTLB2;C5;41.210514;-8.6249732
3976;STCP;0;0;NÓ FREIXIEIRO;1NOFR1;C4;41.219185;-8.6894392
3977;STCP;0;0;E.NACIONAL 107;1ESTN1;C4;41.221418;-8.6869742
3978;STCP;0;0;ANTERO QUENTAL;1RAQ1;C4;41.223381;-8.6844202
3979;STCP;0;0;FARRAPAS;1FARS1;C4;41.222417;-8.6818164
3980;STCP;0;0;PEDRAS RUBRAS;1PRU1;N10;41.238982;-8.669858
3981;STCP;0;0;ASPRELA;1APS3;C9;41.181667;-8.5960266
3982;STCP;0;0;SR. DO CALVÁRIO;SCAL4;C8;41.172023;-8.5592914
3983;STCP;0;0;LABORIM;LBM1;S8;41.107096;-8.6094066
3984;STCP;0;0;FERREIRA DE CASTRO;FCAS1;S8;41.109862;-8.6118568
3985;STCP;0;0;Qta Rosas;QTR2;S8;41.10987;-8.6143083
3986;STCP;0;0;HUMBERTO DELGADO;HMB3;S8;41.109995;-8.6156123
3987;STCP;0;0;FERREIRA DE CASTRO;FCAS2;S8;41.109801;-8.6122063
0;MP;1;MP1;Senhor de Matosinhos;NULL;C3;41.18818954;-8.68516105
1;STCP;1;STCP1;S. ROQUE;SR2;C6;41.16615072;-8.569365092
2;MP;2;MP2;Mercado;NULL;C3;41.1874731;-8.693392192
3;STCP;2;STCP2;TV. FORNO;TVF1;C9;41.19416489;-8.574082896
```

Figure 4.4: Portion of the dataset that has the ids of all station in Andate AFC system

4.4 Temporal Patterns

In this section, it will be described all the data mining techniques and approaches done related with Temporal Patterns. Those techniques are the pre-processing of the data, the clustering methods used to profile the passengers time-wise, more specifically Gaussian Mixture Models(GMM) and K-Means, and the methods Akaike information criterion(AIC) and Bayesian information criterion(BIC) to measure the quality of the results of those clustering methods. Afterwards, we will show how we find the number of new users in the network for a year.

4.4.1 Pre-Processing

Since we were looking for temporal patterns we only used the columns with information related to user and time, more exactly, "cod_cartao" and "datahoravalidacao". These columns are an identifier of each ticket and a date with the hour of transaction time.

The approach that we did, shows the number of transactions in each hour of the day or week, by each passenger, in this case, each "cod_cartao".

ID	H_1	H_2	H_3	H_4	...	H_{21}	H_{22}	H_{23}	H_{24}
id_1	0	0	0	0	...	0	1	1	0
id_2	1	0	0	0	...	0	0	0	2

Table 4.3: Approach for the Day Temporal Data

ID	D_1H_1	...	D_1H_{24}	D_2H_1	...	D_2H_{24}	D_3H_1	...	D_3H_{24}	D_4H_1	...	D_4H_{24}	D_5H_1	...	D_5H_{24}
id_1	1	...	0	1	...	0	1	...	0	1	...	0	0	...	0
id_2	0	...	0	0	...	0	0	...	0	0	...	0	0	...	0

Table 4.4: Approach for the Week Temporal Data

These approaches represented in table4.3 and table4.4 had the purpose of representing for each passenger his temporal profile of utilization of the transport network. The first approach, table4.3, represents the temporal profile of the passengers by day, and the second approach, represents the temporal profile of the passengers by week.

The transformation of the data in the columns "cod_cartao" and "datahoravalidacao" into these approaches had two steps. The first step was to get the values of the hour, minute and day of each transaction. These values were retrieved from the string value in the column "datahoravalidacao". This string "2013-04-16 16:05:36" is an example of a value in the column "datahoravalidacao" were we only would select the values 16 for the hour, 05 for the minutes and 16 for the day. Consequently, transforming that column string into 3 columns, the hour, minute and day columns. Image 4.5 represents the result of this first step transformation.

```
cardId;hour;minute;day
30139615671;16;5;16
20007504249;18;44;22
30135032328;18;44;22
30137770717;18;45;22
30139045277;18;45;22
30140252005;18;45;22
30138820775;16;53;23
20032138873;18;16;23
30141340216;18;16;23
20030219615;19;2;23
20030296936;19;2;23
30139980858;19;2;23
30140524190;18;29;23
30140714349;18;32;23
10002540887;18;32;23
20030244341;18;32;23
30140555872;18;37;23
20007424407;18;46;23
30135779235;18;46;23
30136229617;19;45;23
1394546663957632;20;29;23
1397402817209472;20;29;23
30140080952;6;36;27
30138871288;6;36;27
30135432677;6;38;27
30140811607;6;48;27
30139783791;6;48;27
30140809525;7;6;27
30139824510;7;7;27
30002487481;7;12;27
30136598937;7;20;27
```

Figure 4.5: Portion of the April Temporal Data after the column transformation

The second step was to transform those columns values, represented in figure 4.5, into the data grouping represented in table 4.3 and table 4.4. To accomplish those approaches, in the datasets created in the first step, like the one represented in the figure 4.5, we search for all the transactions that each passenger has done in a day or the days of a week and then verified when those transactions happened. In each passenger profile, we add a value of 1 in the column that corresponds to the transaction hour. The same was done in the second approach, but this time adding a value of 1 in the column that corresponds with the hour and the day of the transaction hour and day. Figure 4.6 shows the result of this last step transformation.

```
id;0;1;2;3;4;5;6;7;8;9;10;11;12;13;14;15;16;17;18;19;20;21;22;23
20007294214;0;0;0;0;0;0;0;0;0;0;0;0;0;2;0;1;0;0;0;0;0;0
20007400647;0;0;0;0;0;0;0;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0
30138458857;0;0;0;0;0;0;0;1;0;0;1;0;0;0;0;0;0;0;0;0;0;0
30136671507;0;0;0;0;0;0;1;0;0;1;0;0;0;0;0;0;0;0;0;0;0;0
80002003086;0;0;0;0;0;0;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
20007254203;0;0;0;0;0;0;2;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
80001474211;0;0;0;0;0;0;1;0;0;0;1;0;0;0;0;0;0;0;0;0;0;0
80001431652;0;0;0;0;0;0;1;0;0;0;1;0;0;0;0;0;0;0;0;0;0;0
10000990678;0;0;0;0;0;0;1;0;1;0;0;0;0;0;0;0;0;0;0;0;0;0
20031310892;0;0;0;0;0;0;2;0;0;0;0;0;1;1;0;0;0;0;0;0;0;0
80001491300;0;0;0;0;0;0;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
80001005933;0;0;0;0;0;0;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
30136340685;0;0;0;0;0;0;0;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0
30003126599;0;0;0;0;0;0;0;0;1;0;0;0;0;0;0;0;0;0;0;0;0;0
80001887754;0;0;0;0;0;0;0;0;1;0;0;0;0;0;0;0;0;0;0;0;0;0
10002741787;0;0;0;0;0;0;0;0;2;0;2;0;0;0;0;0;0;0;0;0;0;0
20007297835;0;0;0;0;0;0;0;0;2;0;2;0;0;0;0;0;0;0;0;0;0;0
30140119082;0;0;0;0;0;0;0;0;1;0;0;0;0;0;0;0;0;0;0;0;0;0
30139762120;0;0;0;0;0;0;0;0;0;1;0;0;0;0;0;0;0;0;0;0;0;0
80001228551;0;0;0;0;0;0;0;0;0;1;0;0;0;0;0;0;0;0;0;0;0;0
30139442014;0;0;0;0;0;0;0;0;2;0;0;2;0;2;0;0;0;0;0;0;0;0
30137582752;0;0;0;0;0;0;0;0;2;2;1;0;0;0;0;0;0;0;0;0;0;0
20007201044;0;0;0;0;0;0;0;0;2;0;0;0;0;0;0;0;0;1;0;0;0;0
30135145399;0;0;0;0;0;0;0;0;1;0;0;1;0;0;0;0;0;0;0;0;0;0
30138837162;0;0;0;0;0;0;0;0;0;1;0;1;0;0;0;0;0;0;0;0;0;0
30139756579;0;0;0;0;0;0;0;0;1;0;1;0;0;0;0;0;0;0;0;0;0;0
30139975144;0;0;0;0;0;0;0;0;1;0;0;0;0;0;0;0;0;0;0;0;0;0
30139760138;0;0;0;0;0;0;0;0;1;3;1;1;0;0;0;0;0;0;0;0;0;0
30139761642;0;0;0;0;0;0;0;0;1;0;1;0;0;0;0;0;0;0;0;0;0;0
30136660620;0;0;0;0;0;0;0;0;0;0;1;2;0;0;0;0;0;0;0;0;0;0
30003327288;0;0;0;0;0;0;0;0;0;0;3;0;0;0;0;0;0;0;0;0;0;0
30135034787;0;0;0;0;0;0;0;0;0;1;0;0;0;0;0;0;0;0;0;0;0;0
30137929851;0;0;0;0;0;0;0;0;0;0;2;0;0;0;0;0;0;0;0;0;0;0
1234537523127424;0;0;0;0;0;0;0;0;0;1;3;0;1;1;0;0;0;0;0;0;0;0
```

Figure 4.6: Portion of Temporal Data day approach of January 2, 2013

4.4.2 Gaussian Mixture Models

Now that we have the data in the desired format, we can run a Gaussian Mixture Models algorithm[18] on the data.

The Gaussian Mixture Models algorithm has various parameters to tune. In our experiences, we will only use the "n_components" and "covariance_type" parameters. These parameters are, respectively, the number of Gaussians that we will try to fit in the data(the number of clusters), and the type of covariance parameters to use.

4.4.2.1 Akaike Information Criterion And Bayesian Information Criterion

Akaike Information Criterion(AIC) And Bayesian Information Criterion(BIC) are used to measure the quality of the model. Both these criteria use a penalty term because of the overfitting problem that occurs when adding more parameters. This penalty term is larger in BIC. These criteria are defined as follows:

$$AIC = 2\kappa - 2\ln(\hat{L}) \tag{4.1}$$

- κ = the number of parameters of the model
- $\hat{L}$ = the maximum value of the likelihood function for the model

$$BIC = \kappa\ln(n) - 2\ln(\hat{L}) \tag{4.2}$$

- κ = the number of parameters of the model
- $\hat{L}$ = the maximum value of the likelihood function for the model
- n = the sample size

For both criteria, the best model is the one with the lower result. As said before, both criteria have a penalty term. In the AIC, the penalty term is 2κ part, so it penalizes by the number of parameters. In the BIC, the penalty term is $\kappa\ln(n)$ part, so it penalizes by the number of parameters and by the sample size.

We used these criteria to find the best tune for the GMM "n_components" and "covariance_type" parameters. Our approach was to find the best tune for the temporal data of a day, of a week and a weekday.

4.4.2.2 Temporal Data of a Day

The finding of the best tune for the GMM parameters for the temporal data of a day was done in the following way. We calculated the AIC and BIC for different values in the "n_components" with the same value in "covariance_type" and plot it. Then we repeated this process for each value that we could put in the "covariance_type" parameter. The values of that parameter were "full", "tied", "diag" and "spherical". Afterwards, we plotted each AIC and BIC by the "covariance_type" values. Here is an example of the plots to tune the parameters for a day(January 7, 2013):

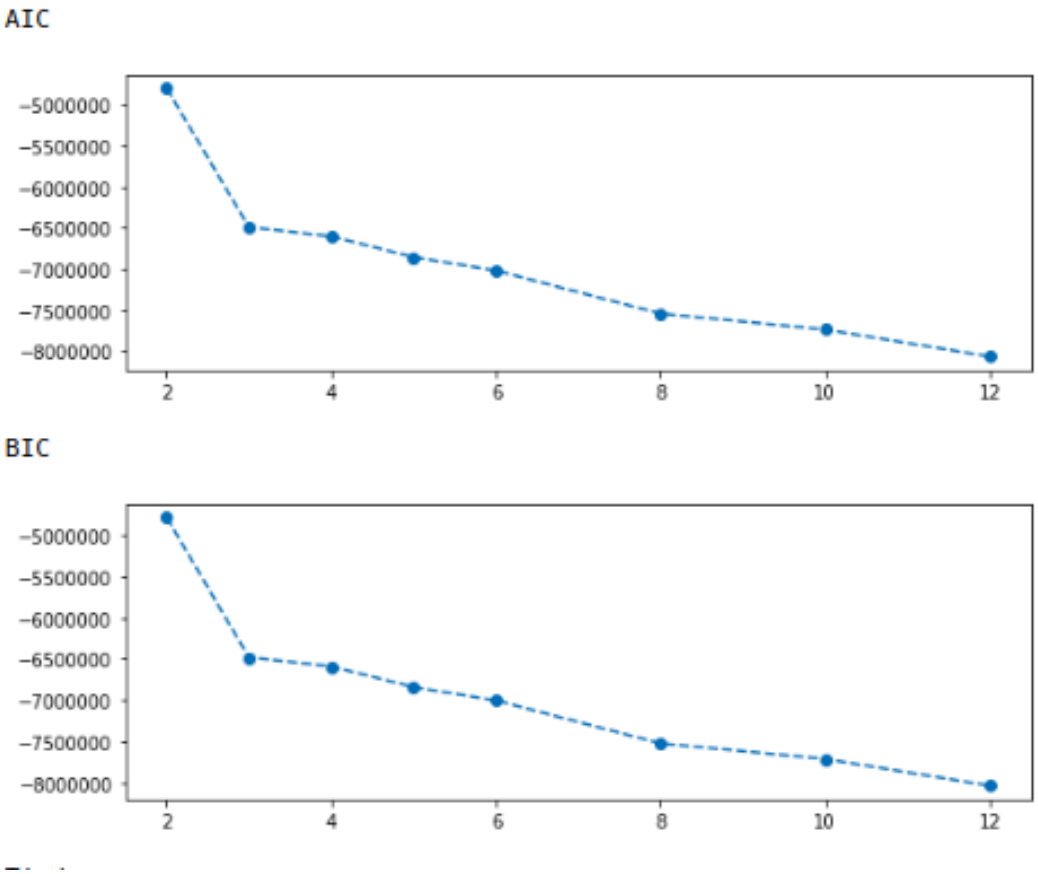


Figure 4.7: AIC and BIC plots for value "full" of the parameter "covariance_type"

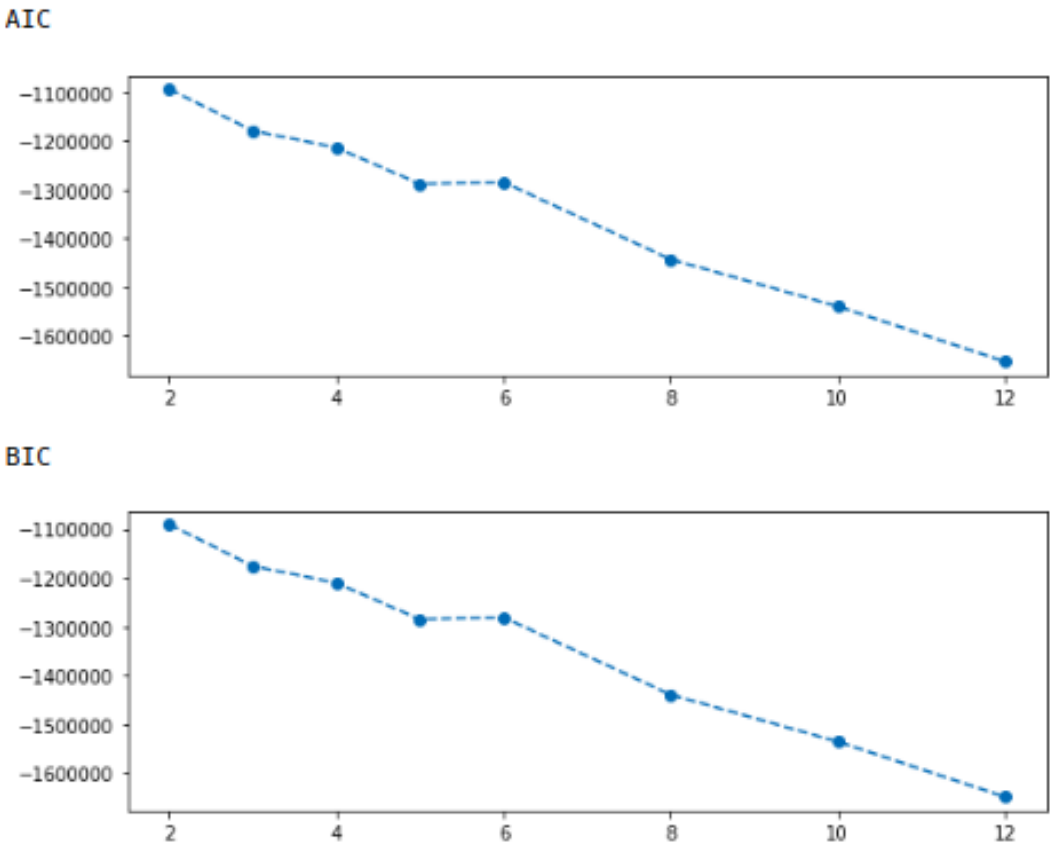


Figure 4.8: AIC and BIC plots for value "tied" of the parameter "covariance_type"

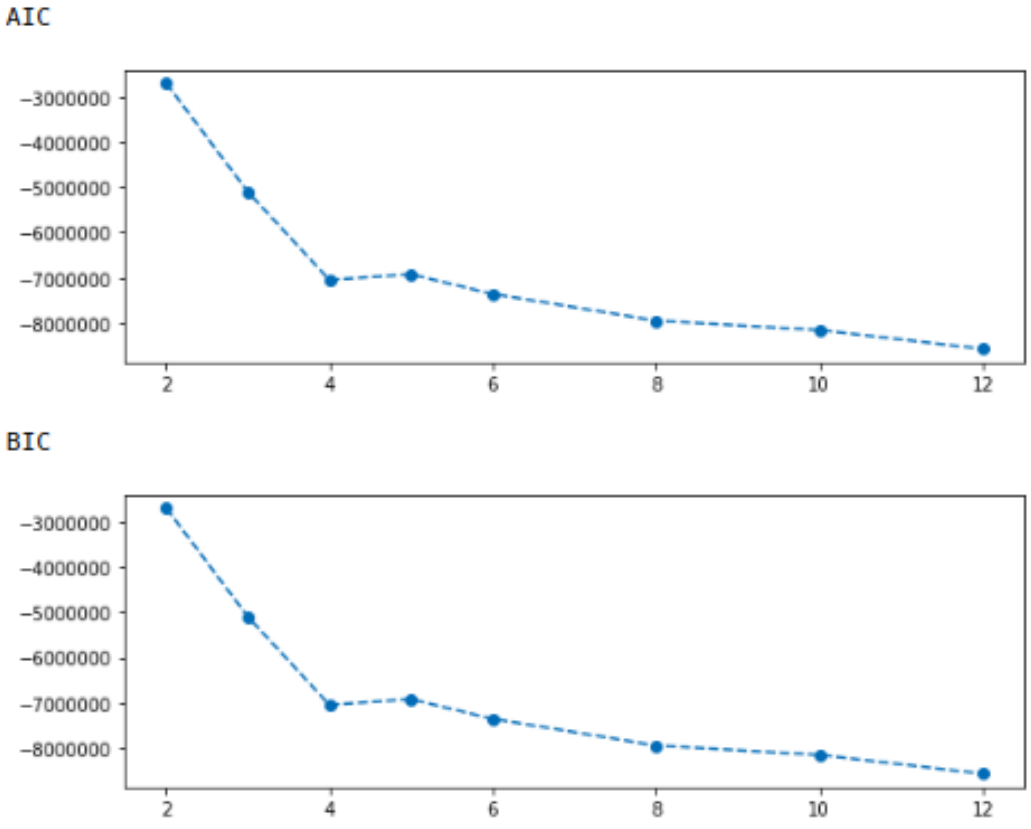


Figure 4.9: AIC and BIC plots for value "diag" of the parameter "covariance_type"

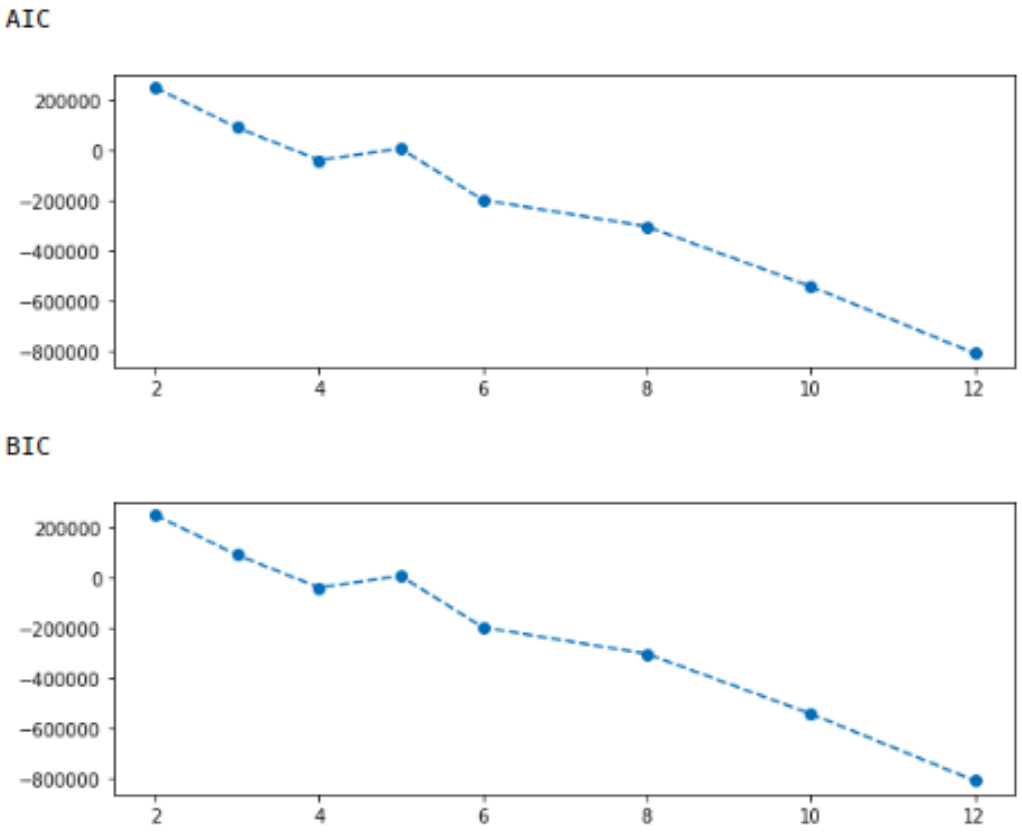


Figure 4.10: AIC and BIC plots for value "spherical" of the parameter "covariance_type"

The results show that the values "full" and "diag" for the "covariance_type" parameter have the lowest AIC and BIC values. This means that they will probably represent a better model. Afterwards, we decided that for the parameter "n_components" we would use the values "3" or "4". These values were chosen because from these values on that were introduced in the "n_components" parameters, the difference between the AIC values of each value introduced in the "n_components" parameters becomes fewer and fewer, as can be seen in the Figures 4.7 and 4.9. The same happens in the BIC values. The values of AIC and BIC get lower, but they are probably a sign of overfitting.

4.4.2.3 Temporal Data for a week

The finding of the best tune for the GMM parameters for the temporal data of the week was done in the same way as it was done for the day. The use of the AIC and BIC criteria to find the best values for the "n_components" and "covariance_type" parameters. Here we gonna use the week data from January 7 to January 11, 2013, as an example to show how the plots were used to tune the parameters for the week:

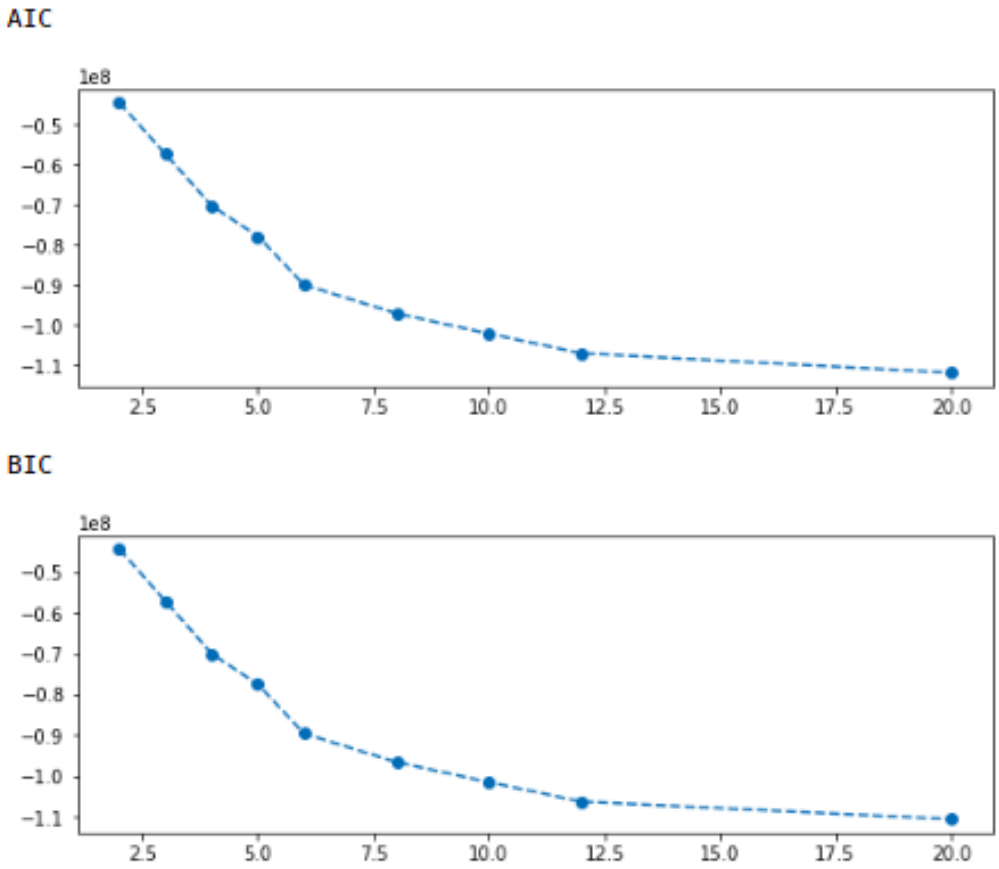


Figure 4.11: AIC and BIC plots for value "full" of the parameter "covariance_type"

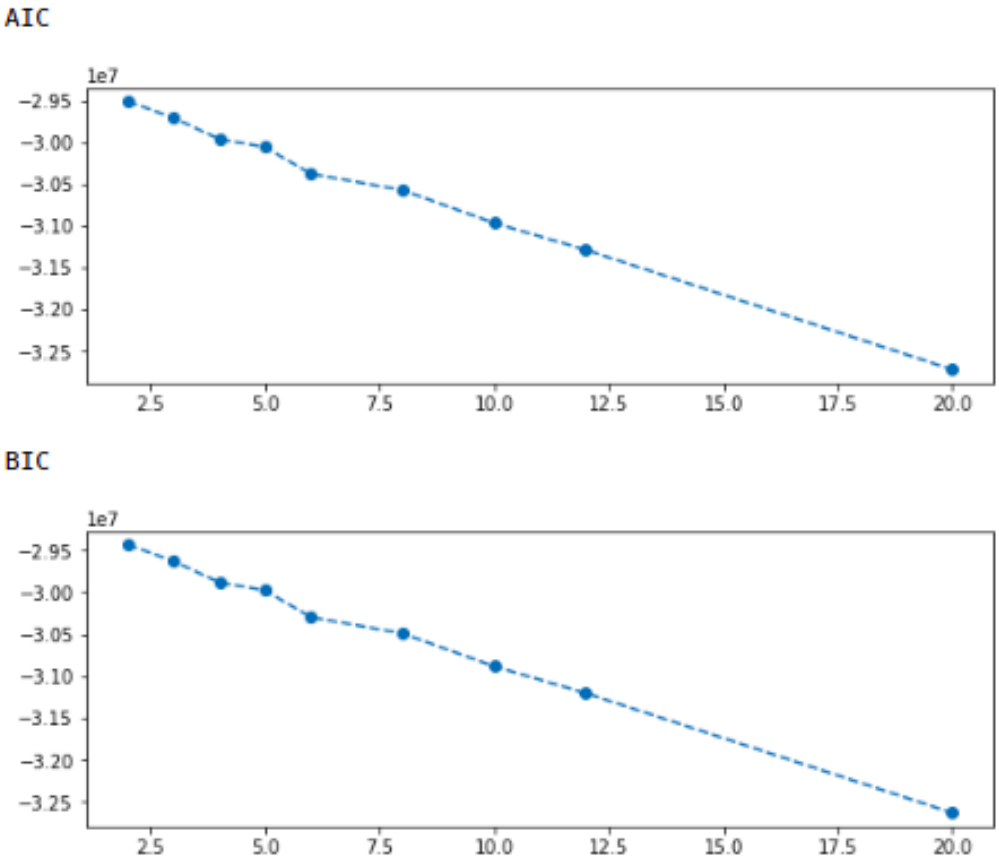


Figure 4.12: AIC and BIC plots for value "tied" of the parameter "covariance_type"

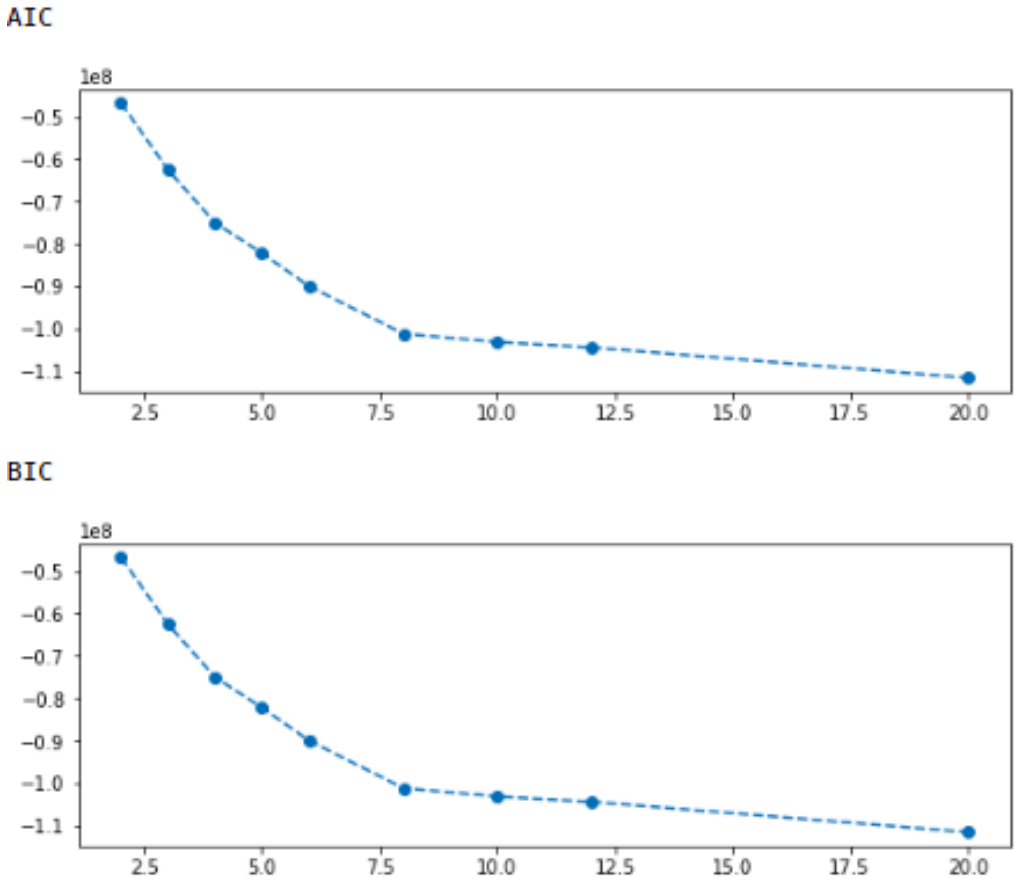


Figure 4.13: AIC and BIC plots for value "diag" of the parameter "covariance_type"

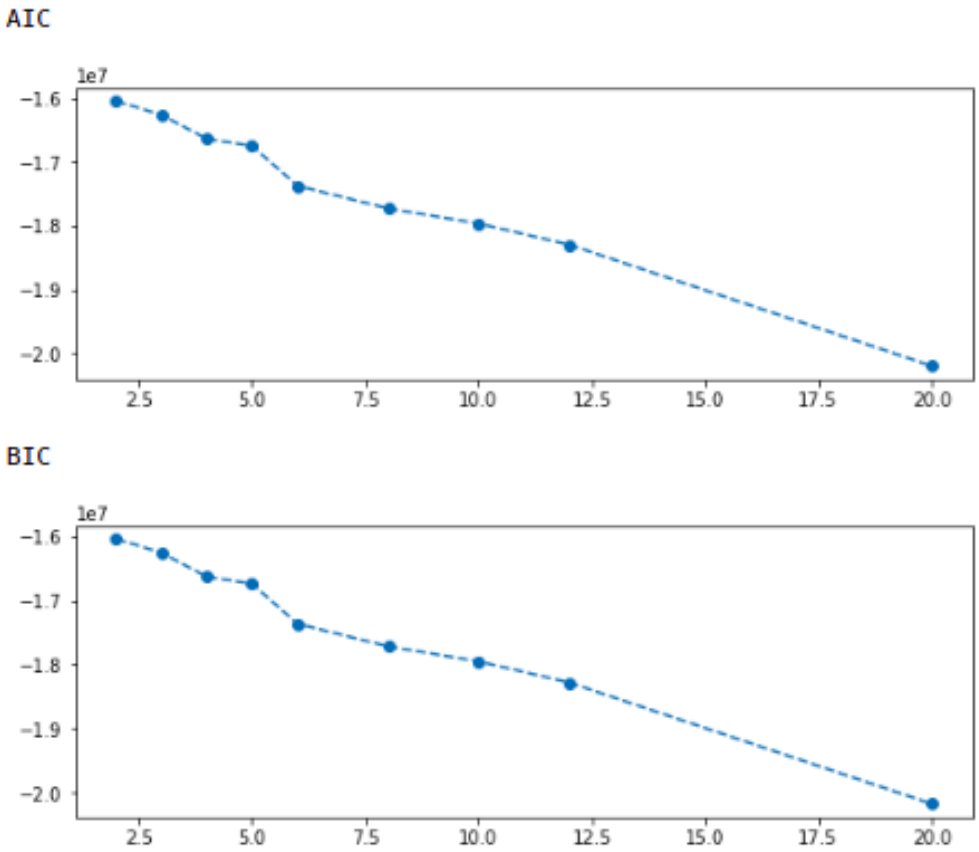


Figure 4.14: AIC and BIC plots for value "spherical" of the parameter "covariance_type"

The results show, once again, that the values "full" and "diag" for the "covariance_type" parameter have the lowest AIC and BIC values. The values that we decided to use for the "n_components" parameter were "3", "4" and "6". These values were chosen because they are the interval of values that still has a significant difference between their AIC values, and also a significant difference between their BIC values, as can be seen in the Figures 4.7 and 4.9 . From these values on those differences become fewer and fewer.

4.4.2.4 Results

In this section, we are only describing one example of the result of the GMM when utilized in the temporal data for a day and a week. The result of the GMM algorithm is plotted by clusters. To ease the understanding of the temporal pattern, we plot all the cluster transactions that occur in each hour. The following two examples represent the results from temporal data of a day and a week.

This first example represents the result of the GMM algorithm with the parameters "covariance_type" and "n_components" set as "diag" and "4", respectively, in the temporal data of day January 7, 2013.

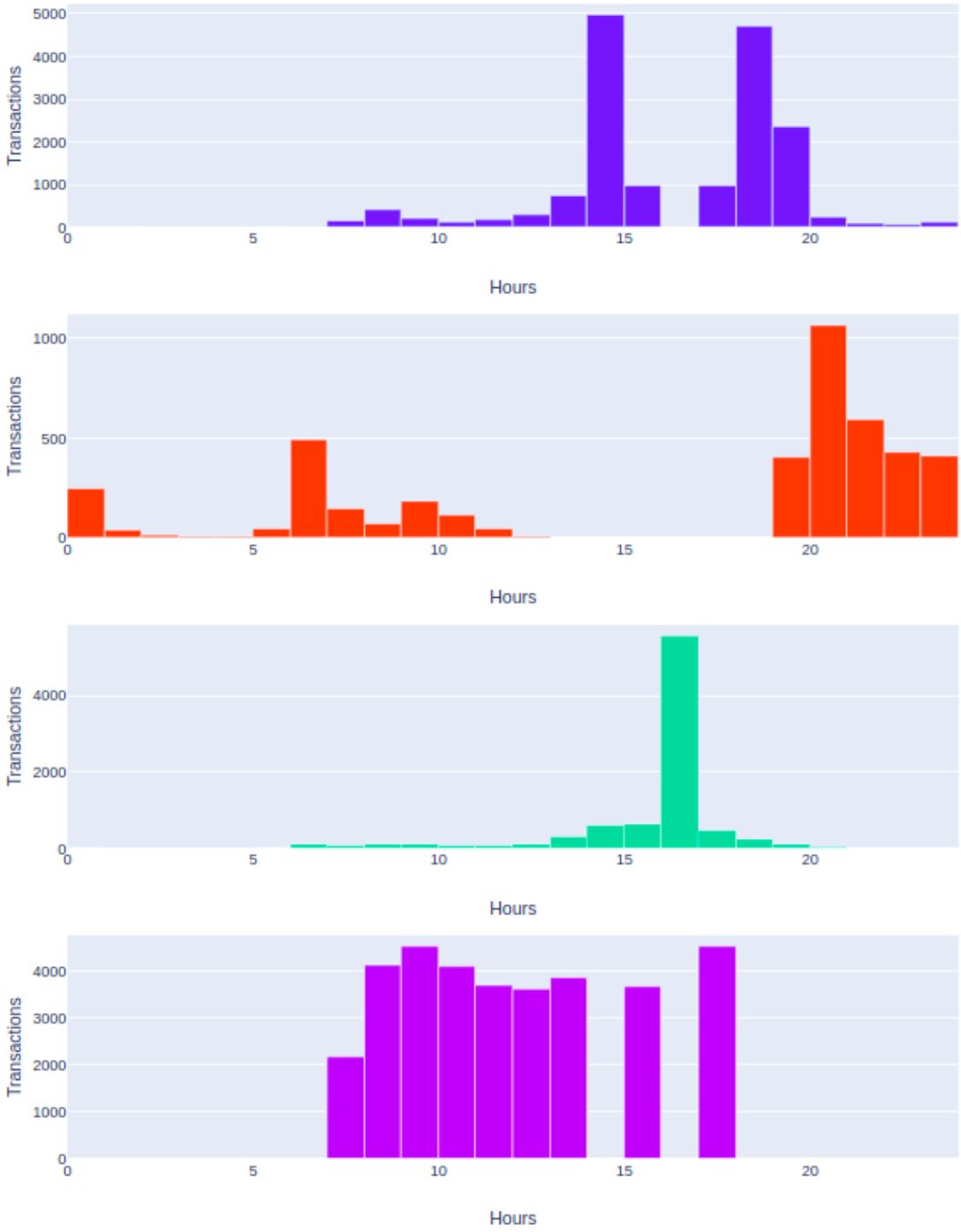


Figure 4.15: Example of GMM Algorithm result of January 7, 2013

This second example represents the result of the GMM algorithm with the parameters "covariance_type" and "n_components" set as "full" and "6", respectively, in the temporal data of week 7-11 January, 2013.

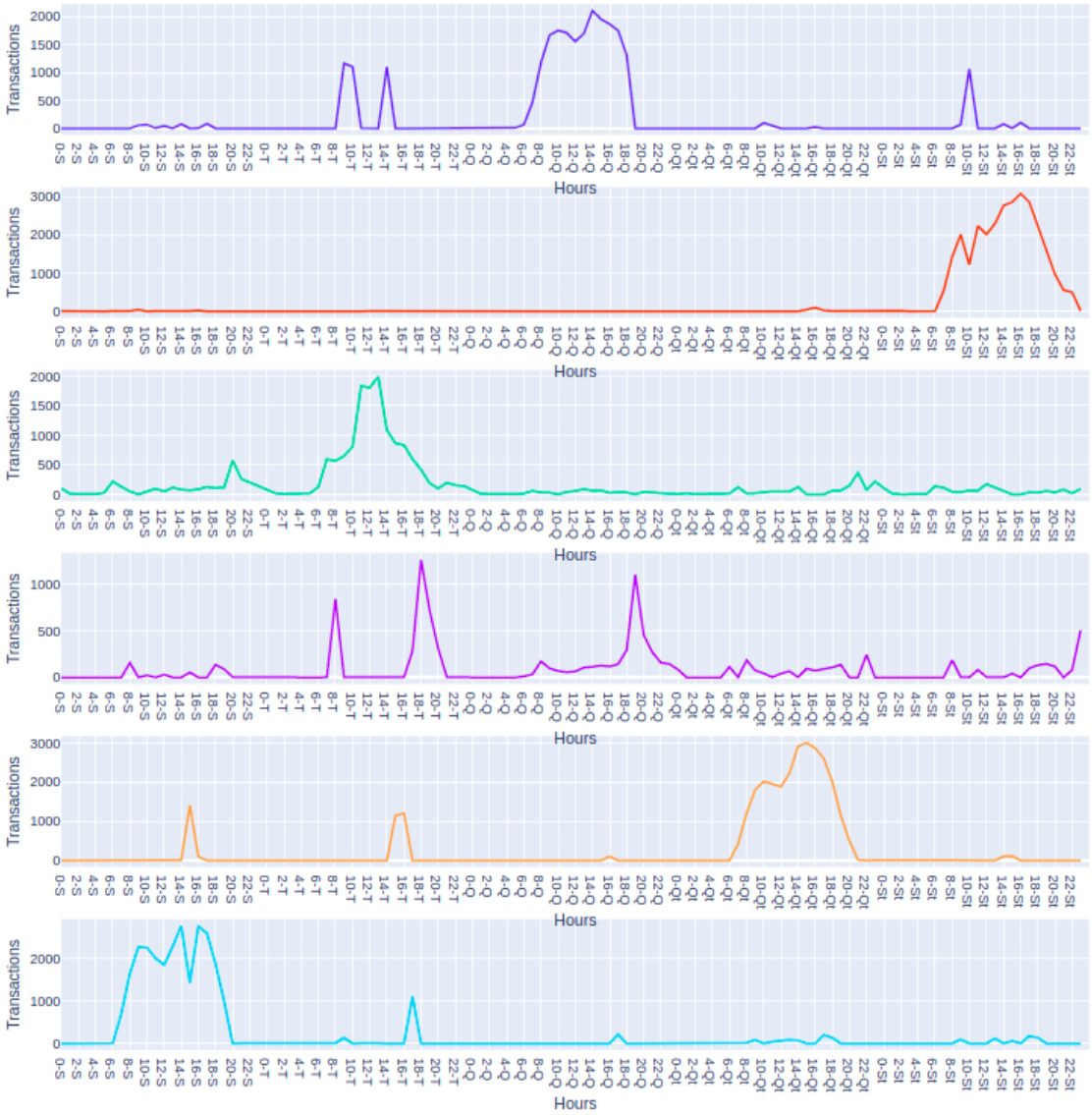


Figure 4.16: Example of GMM Algorithm result of January 7 to 11 2013

4.4.3 New Passengers

This approach that will be described here has the objective of finding the number of new passengers each day. This will help us to understand how many new ticket are being used, and if that value of new tickets varies or not, in each day.

The approach to finding the new passengers of each day is achieved by comparing the old ids(ids that already exists in the system) with the ids of that day. The values of new passengers in January and February are artificially high because the old ids pool start empty in January and for that reason in February we are still catching new id's that aren't new passengers. We can't do anything about it since we only have data from January. We also made a ratio comparing the new

ticket with the tickets used, in each day.

$$RatioOfNewTickets = \frac{NumberOfNewTicketsIds}{NumberOfAllTicketsUsed} \tag{4.3}$$

Figure 4.17 will show two months of approach.

Day	New Tickets	Day Ratio of New Tickets
1	25237	1.000000
2	52265	0.928264
3	42284	0.713016
4	33555	0.572211
5	22128	0.595223
6	11115	0.501263
7	27059	0.493498
8	22079	0.425054
9	19671	0.382110
10	19304	0.355631
11	19201	0.340407
12	10902	0.371955
13	6700	0.336650
14	16554	0.320752
15	14593	0.292550
16	12022	0.272651
17	12924	0.262001
18	11903	0.255243
19	7389	0.300537
20	4131	0.262920
21	13132	0.261573
22	10208	0.232778
23	10090	0.223621
24	12138	0.231509
25	12419	0.229323
26	9371	0.282480
27	3934	0.232026
28	13573	0.245013
29	11278	0.217332
30	11360	0.214364
31	11505	0.207608

(a) Results of January new passengers

Day	New Tickets	Day Ratio of New Tickets
1	11414	0.174153
2	9195	0.151770
3	8821	0.144122
4	8438	0.140755
5	8589	0.143499
6	8081	0.206712
7	5252	0.185872
8	8260	0.148925
9	8157	0.144259
10	8439	0.140185
11	8465	0.141060
12	8374	0.140463
13	6271	0.174709
14	4822	0.183102
15	8804	0.154915
16	8241	0.142874
17	8423	0.142353
18	8515	0.145270
19	8440	0.144915
20	6708	0.181858
21	4839	0.190280
22	8255	0.153684
23	8157	0.144420
24	8122	0.141343
25	8055	0.142771
26	7928	0.138941
27	5994	0.176502
28	6393	0.227485
29	7729	0.145663
30	7836	0.141044
31	8091	0.144464

(b) Results of July new passengers

Figure 4.17: Examples of Results of New Passengers Approach

4.5 Spatial Patterns

In this section, we will describe all the data mining techniques and approaches related with Spatial Patterns. Those techniques are the data pre-processing of the data, the DBSCAN clustering method and its respective tuning, that was used to find the places with more affluence in the network. Also, in this section, we will discuss the method for finding the type of ticket most used.

4.5.1 Pre-Processing

For the research on Spatial Patterns we needed the raw data columns with information about where the passengers were in spatially. For this purpose, we selected the columns "cod_cartao", "Paragem", "datahoravalidacao", "Zona", "lat" and "long". We also renamed some of the columns for better understanding. We renamed "cod_cartao" for "id", "Paragem" for "station" and "datahoravalidacao" for "timestamp" as Figure 4.18 shows.

```
id;station;zona;timestamp;lat;long
30139615671;3956;2013-04-16 16:05:36;C2;41.163776;-8.686086999999999
20007504249;3956;2013-04-22 18:44:35;C2;41.163776;-8.686086999999999
30135032328;3956;2013-04-22 18:44:55;C2;41.163776;-8.686086999999999
30137770717;3956;2013-04-22 18:45:03;C2;41.163776;-8.686086999999999
30139045277;3956;2013-04-22 18:45:04;C2;41.163776;-8.686086999999999
30140252005;3956;2013-04-22 18:45:09;C2;41.163776;-8.686086999999999
30138820775;3956;2013-04-23 16:53:13;C2;41.163776;-8.686086999999999
20032138873;3956;2013-04-23 18:16:07;C2;41.163776;-8.686086999999999
30141340216;3956;2013-04-23 18:16:11;C2;41.163776;-8.686086999999999
20030219615;3956;2013-04-23 19:02:41;C2;41.163776;-8.686086999999999
20030296936;3956;2013-04-23 19:02:49;C2;41.163776;-8.686086999999999
30139980858;3956;2013-04-23 19:02:50;C2;41.163776;-8.686086999999999
30140524190;3956;2013-04-23 18:29:15;C2;41.163776;-8.686086999999999
30140714349;3956;2013-04-23 18:32:10;C2;41.163776;-8.686086999999999
10002540887;3956;2013-04-23 18:32:41;C2;41.163776;-8.686086999999999
20030244341;3956;2013-04-23 18:32:58;C2;41.163776;-8.686086999999999
30140555872;3956;2013-04-23 18:37:45;C2;41.163776;-8.686086999999999
20007424407;3956;2013-04-23 18:46:30;C2;41.163776;-8.686086999999999
30135779235;3956;2013-04-23 18:46:29;C2;41.163776;-8.686086999999999
30136229617;3956;2013-04-23 19:45:13;C2;41.163776;-8.686086999999999
30140080952;3956;2013-04-27 06:36:36;C2;41.163776;-8.686086999999999
30138871288;3956;2013-04-27 06:36:43;C2;41.163776;-8.686086999999999
30135432677;3956;2013-04-27 06:38:22;C2;41.163776;-8.686086999999999
30140811607;3956;2013-04-27 06:48:19;C2;41.163776;-8.686086999999999
30139783791;3956;2013-04-27 06:48:22;C2;41.163776;-8.686086999999999
30140809525;3956;2013-04-27 07:06:57;C2;41.163776;-8.686086999999999
30139824510;3956;2013-04-27 07:07:45;C2;41.163776;-8.686086999999999
30002487481;3956;2013-04-27 07:12:32;C2;41.163776;-8.686086999999999
30136598937;3956;2013-04-27 07:20:06;C2;41.163776;-8.686086999999999
30140253188;3956;2013-04-29 07:26:30;C2;41.163776;-8.686086999999999
20031300533;3956;2013-04-29 07:32:49;C2;41.163776;-8.686086999999999
30136619496;3956;2013-04-29 07:37:18;C2;41.163776;-8.686086999999999
```

Figure 4.18: Columns Selected for Spacial data

Afterwards, we separated the column "timestamp" into three columns. The columns "hour", "minute" and "day". The objective of this transformation is to make the information that the column "timestamp" holds easy to work with. Figure 4.19 shows the data format that we will work with, in this section.

```
id;statio;zona;lat;long;hour;min;day
30139615671;3956;C2;41.163776;-8.686086999999999;16;5;16
20007504249;3956;C2;41.163776;-8.686086999999999;18;44;22
30135032328;3956;C2;41.163776;-8.686086999999999;18;44;22
30137770717;3956;C2;41.163776;-8.686086999999999;18;45;22
30139045277;3956;C2;41.163776;-8.686086999999999;18;45;22
30140252005;3956;C2;41.163776;-8.686086999999999;18;45;22
30138820775;3956;C2;41.163776;-8.686086999999999;16;53;23
20032138873;3956;C2;41.163776;-8.686086999999999;18;16;23
30141340216;3956;C2;41.163776;-8.686086999999999;18;16;23
20030219615;3956;C2;41.163776;-8.686086999999999;19;2;23
20030296936;3956;C2;41.163776;-8.686086999999999;19;2;23
30139980858;3956;C2;41.163776;-8.686086999999999;19;2;23
30140524190;3956;C2;41.163776;-8.686086999999999;18;29;23
30140714349;3956;C2;41.163776;-8.686086999999999;18;32;23
10002540887;3956;C2;41.163776;-8.686086999999999;18;32;23
20030244341;3956;C2;41.163776;-8.686086999999999;18;32;23
30140555872;3956;C2;41.163776;-8.686086999999999;18;37;23
20007424407;3956;C2;41.163776;-8.686086999999999;18;46;23
30135779235;3956;C2;41.163776;-8.686086999999999;18;46;23
30136229617;3956;C2;41.163776;-8.686086999999999;19;45;23
30140080952;3956;C2;41.163776;-8.686086999999999;6;36;27
30138871288;3956;C2;41.163776;-8.686086999999999;6;36;27
30135432677;3956;C2;41.163776;-8.686086999999999;6;38;27
30140811607;3956;C2;41.163776;-8.686086999999999;6;48;27
30139783791;3956;C2;41.163776;-8.686086999999999;6;48;27
30140809525;3956;C2;41.163776;-8.686086999999999;7;6;27
30139824510;3956;C2;41.163776;-8.686086999999999;7;7;27
30002487481;3956;C2;41.163776;-8.686086999999999;7;12;27
30136598937;3956;C2;41.163776;-8.686086999999999;7;20;27
30140253188;3956;C2;41.163776;-8.686086999999999;7;26;29
20031300533;3956;C2;41.163776;-8.686086999999999;7;32;29
30136619496;3956;C2;41.163776;-8.686086999999999;7;37;29
10002613156;3956;C2;41.163776;-8.686086999999999;7;37;29
```

Figure 4.19: Spatial Data Format

For finding the most used ticket type we used a similar dataset. The only difference was that we added a column, "zonas", that holds the information about the ticket types. So, the dataset was as the following Figure 4.20.

	id	station	ticketType	zona	lat	long	hour	min	day
0	30138874701	3956	Z2	C2	41.163776	-8.686087	7	12	18
1	30139476831	3956	Z2	C2	41.163776	-8.686087	13	8	8
2	30139038925	3956	Z2	C2	41.163776	-8.686087	13	23	8
3	30002922101	3956	Z2	C2	41.163776	-8.686087	13	29	8
4	30136583083	3956	Z3	C2	41.163776	-8.686087	13	43	8
5	30139772827	3956	Z3	C2	41.163776	-8.686087	13	1	3
6	10000755837	3956	Z3	C2	41.163776	-8.686087	13	1	3
7	30137462137	3956	Z2	C2	41.163776	-8.686087	13	1	3
8	30138448357	3956	Z2	C2	41.163776	-8.686087	13	1	3
9	30003193899	3956	Z2	C2	41.163776	-8.686087	19	3	19
10	30139358120	3956	Z4	C2	41.163776	-8.686087	16	23	3
11	20031333061	3956	Rede Geral(ABC)	C2	41.163776	-8.686087	16	23	3
12	30134301076	3956	Z2	C2	41.163776	-8.686087	6	27	23
13	30136739411	3956	Z2	C2	41.163776	-8.686087	6	21	23
14	30138826334	3956	Z2	C2	41.163776	-8.686087	8	1	22
15	30139760178	3956	Z3	C2	41.163776	-8.686087	8	3	22
16	30141329517	3956	Z3	C2	41.163776	-8.686087	8	8	22
17	30139778070	3956	Z2	C2	41.163776	-8.686087	8	8	22
18	20031180563	3956	Z2	C2	41.163776	-8.686087	8	9	22
19	30136784131	3956	Z2	C2	41.163776	-8.686087	8	9	22
20	30005393965	3956	Z3	C2	41.163776	-8.686087	8	9	22
21	30004353711	3956	Z3	C2	41.163776	-8.686087	8	13	22
22	30005686981	3956	Z3	C2	41.163776	-8.686087	8	14	22
23	30134793396	3956	Z4	C2	41.163776	-8.686087	8	16	22
24	30139814785	3956	Z2	C2	41.163776	-8.686087	8	17	22
25	30004603654	3956	Z2	C2	41.163776	-8.686087	8	42	22
26	30139740638	3956	Z2	C2	41.163776	-8.686087	18	3	20
27	20032169069	3956	Z5	C2	41.163776	-8.686087	17	15	19
28	30139787581	3956	Z2	C2	41.163776	-8.686087	17	20	19
29	30140112910	3956	Z3	C2	41.163776	-8.686087	17	20	19

Figure 4.20: Ticket Type Dataset

4.5.2 DBSCAN

We used the DBSCAN algorithm to search for the most common areas of the transport network. Since DBSCAN is a density-based algorithm. Its characteristics were a benefit for the purpose that we had. That purpose, as said before was to discover the areas that are the most commonly used in the network. For this purpose, we used in DBSCAN the columns "lat" and "long" from the dataset represented in Figure 4.19.

The DBSCAN algorithm has some parameters that need to be defined. Two of those parameters are the "eps" and the "min_samples" and we tuned them to make the algorithm have a better result. So, we used the data, from January 7, 2013, as an exampleof the tuning of these parameters. The data is composed of the columns "lat" and "long", as said before, and has 109255 entries. The tuning of these two parameters was done separately, mainly to see the impact of each parameter on the results. The "min_samples" is the parameter that holds the minimal value of points that need to be "closer" to make a cluster. So, these parameters are useful to only choose areas that have a minimal value of validations. The "eps" is the parameter that holds the value of the radius of the area around each point. The larger the "eps" the longer can be the distance between points to be in the same cluster, the lower the "eps" the shorter can be the distance between points to be in the same cluster.

MIN_PTS	Clusters	NoisePoints
25	345	3376
50	227	7794
100	140	14850
250	66	27563
500	37	38753
1000	20	51351
1250	15	58087
1500	12	62509
2000	9	68063
5000	1	96587

(a) "mim_samples" Tuning Results("eps" = 0.002)

EPS	Clusters	NoisePoints
0,0005	159	26485
0,001	153	22631
0,0015	148	18700
0,002	140	14850
0,0025	120	10281
0,003	79	7038
0,0035	49	5387
0,004	35	4417
0,0045	25	3583
0,005	22	3148
0,01	8	212

(b) "eps" Tuning Results("min_samples" = 100)

Figure 4.21: Tuning of the parameters "min_samples" and "eps"

The Figures in 4.21 show the different values assign to each parameter and their results. The "min_samples" value that we chose was 1000. This value is high and almost, half of the points in the dataset was not selected to any cluster. This value also guarantees that we have at least 1000 validations in the clusters, and that is the reason why we chose it 1000. We need a high value, or at least, a significant value of transactions to say that an area in the network is widely used. The

"eps" value that we choose was 0.002. We find that this value was the most appropriated in spacial terms. The values that we tried above make the clusters big and unsuitable for the case because zones that are different locations by themselves then became a bigger location. The vice-versa occurs when the value is below 0.002. the zones that we know spatially start to being broke into various zones. So to know the most common areas of the network we used the "min_samples" as 1000 and the "eps" as 0.002.

4.5.2.1 Results

In this section, we will show how we represented the results of the most commonly used areas of the network. After the DBSCAN result, we plot every cluster into a map of the city of Porto. This way we can understand and visualize the areas of the clusters. Figure 4.22 represents a result of January 7, 2013.

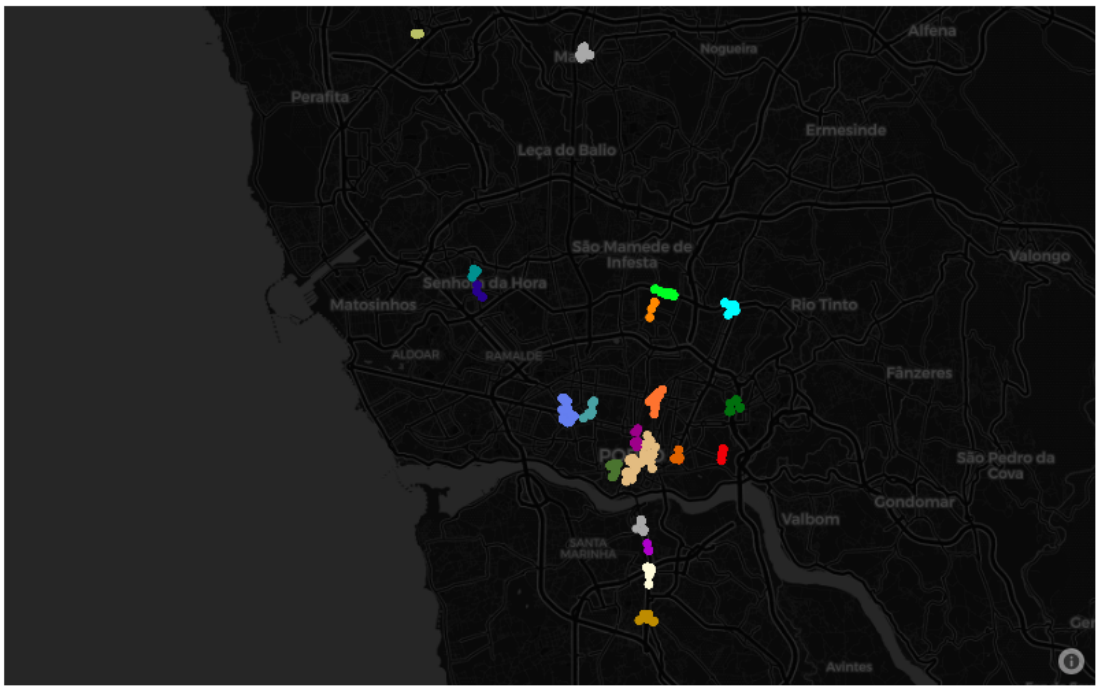


Figure 4.22: DBSCAN Result of January 7, 2013

4.5.3 Ticket Type

The ticket type study had the objective of understanding which of the ticket types has more use in the network by the non-frequent passengers. For this purpose, we used the datasets in the format that is shown in Figure 4.20. We made this analysis using the following two steps.

First, we verified all the unique values that the column "ticketType" had and put them on a list. This results in a list like the following.

```
['Z2', 'Z3', 'Z32', 'T2', 'T1', 'Z4', 'Z5', 'Z6', 'Z7', 'Z8', 'T3', 'Z1']
```

Figure 4.23: Ticket Types Example List

Afterwards, we count all the transactions that had each one of these types. We also divided each value by the number of transactions that the dataset holds. This way we can know the percentage of each type of ticket. The following table is an example of a result.

Ticket Type	Sum of Transaction	Percentage of All Transactions
'Z2'	15492	71.52
'Z3'	3087	14.25
'T1'	834	3.85
'Z4'	1601	7.39
'Z6'	239	1.10
'Z5'	146	0.67
'T2'	176	0.81
'Z32'	53	0.24
'Z8'	2	0.009
'Z7'	2	0.009
'T3'	29	0.13

Table 4.5: Example of a Result from February 4 of The Ticket Type Methodology

This methodology can be done in different types of datasets. By this we mean that we could use a dataset from a day, a week, a month, a year, one time users, two time users, etc. We can choose which type of dataset to put in temporary terms, or usage terms. We just need to make sure that respects the format in Figure 4.20.

4.6 Spatio-Temporal Patterns

In this section, it will be described all the data mining techniques and approaches done related to Spatio−Temporal Patterns. Those techniques are the data pre-processing, the K-Means clustering method and its respective tuning, that was used to find the Spatio-Temporal patterns existent in the network, and an approach to trip-chain.

4.6.1 Pre-Processing

Since we already had in the dataset for the Spatial Patterns all columns that are related to space and time, we used that dataset again, or to make clear, the same dataset format, composed by the columns "id", "station", "Zona", "lat", "long", "hour", "minute" and "day" as Figure 4.19

4.6.2 K-Means

We utilized the K-Means algorithm to Cluster data in spatially and temporal terms. Our objective is to show profiles of space and time of the transport network utilization. Combining the ideas from Temporal and Spacial patterns we can get a profile that has the results from the Temporal patterns and Spacial patterns in one. Since we are clustering based in two different terms, space and time, we opted by using the K-Means. K-Means seams always a safe choice in clustering.

So, we tuned the most important parameter on K-Means, that is the number of clusters. In this algorithm, we need to set the number of clusters that it will search for in the data and that is achieved with this parameter. Consequently, we needed to ensure that the number of clusters that we choose will give us any good results. To ensure that, we use the Elbow Method. This method consists of plotting the sum of squared distance between the data points and their cluster centroid. Therefore, a lower value means better clusters, that means data points of each cluster closer to each other. Whereas a higher value means worse clusters, which means data points of each cluster more disperse. By using this method we can ensure that each cluster has similar data points in their cluster, or in another way, the data points of the cluster are closer to each other. Here is an example of the Elbow Method result in a day data:

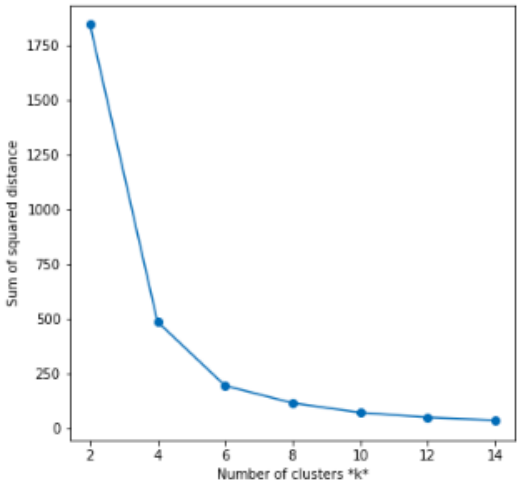


Figure 4.24: Elbow Method Result from January 7

We decided to use 4 and 6 for the number of clusters based on the Elbow method results, Fig.4.24. In this cluster, values are where the plot starts to be flatter. This means the sum squared distance between the data points and the cluster centroid hasn't changed that much. Therefore, we already have a good number of clusters and if we add more clusters we are probably dividing a cluster in half, or another cluster without an impact for the result.

In this K-Means approach to profile the network, in terms of Space and Time, we used only the columns "station" and "hour". Since the station column is an identifier of each station, we used it instead of the latitude and longitude values of each station because the stations are our "network space". In terms of time, we only used hours since we only use the data of a day.

A problem that the algorithm could have was the giving of more weight to the column "station" values since the interval of values existent goes from 0 to 4074. Whereas the column "hour" interval of values goes from 0 to 23. For this big difference in an interval of values, and taking into account that we wanted both columns to have the same importance, we normalize the values and then use the K-Means algorithm on this data. The following figure 4.25 will show how the values were before and after the normalization.

0	2774	6
1	1775	7
2	1235	7
3	2152	7
4	3722	7
5	1420	7
6	1564	7
7	1579	7
8	1705	8
9	1212	8
10	1212	8
11	1231	8
12	1379	8
13	3722	8
14	3722	8
15	1420	8
16	1420	8
17	1564	8
18	1564	8
19	1564	8
20	1579	9
21	1579	9
22	2681	9
23	3055	9
24	1653	9
25	1653	9
26	1493	9
27	1212	9
28	1379	9
29	1379	9

(a) Values before normalization

0	0.999998	0.002163
1	0.999992	0.003944
2	0.999984	0.005668
3	0.999995	0.003253
4	0.999998	0.001881
5	0.999988	0.004930
6	0.999990	0.004476
7	0.999990	0.004433
8	0.999989	0.004692
9	0.999978	0.006601
10	0.999978	0.006601
11	0.999979	0.006499
12	0.999983	0.005801
13	0.999998	0.002149
14	0.999998	0.002149
15	0.999984	0.005634
16	0.999984	0.005634
17	0.999987	0.005115
18	0.999987	0.005115
19	0.999987	0.005115
20	0.999984	0.005700
21	0.999984	0.005700
22	0.999994	0.003357
23	0.999996	0.002946
24	0.999985	0.005445
25	0.999985	0.005445
26	0.999982	0.006028
27	0.999972	0.007426
28	0.999979	0.006526
29	0.999979	0.006526

(b) Values after normalization

Figure 4.25: Normalization of the "station" and "hour" values

4.6.2.1 Results

After all those steps that we talked before that we had to do for the K-Means have better results, we represented the results. The way we used to represent the results is a combination of Figures 4.15 and 4.22. This combination represents each cluster profile in terms of Space and Time. The cluster is in the same order in the temporal plots and the spacial map. The colors can differ because they are random.

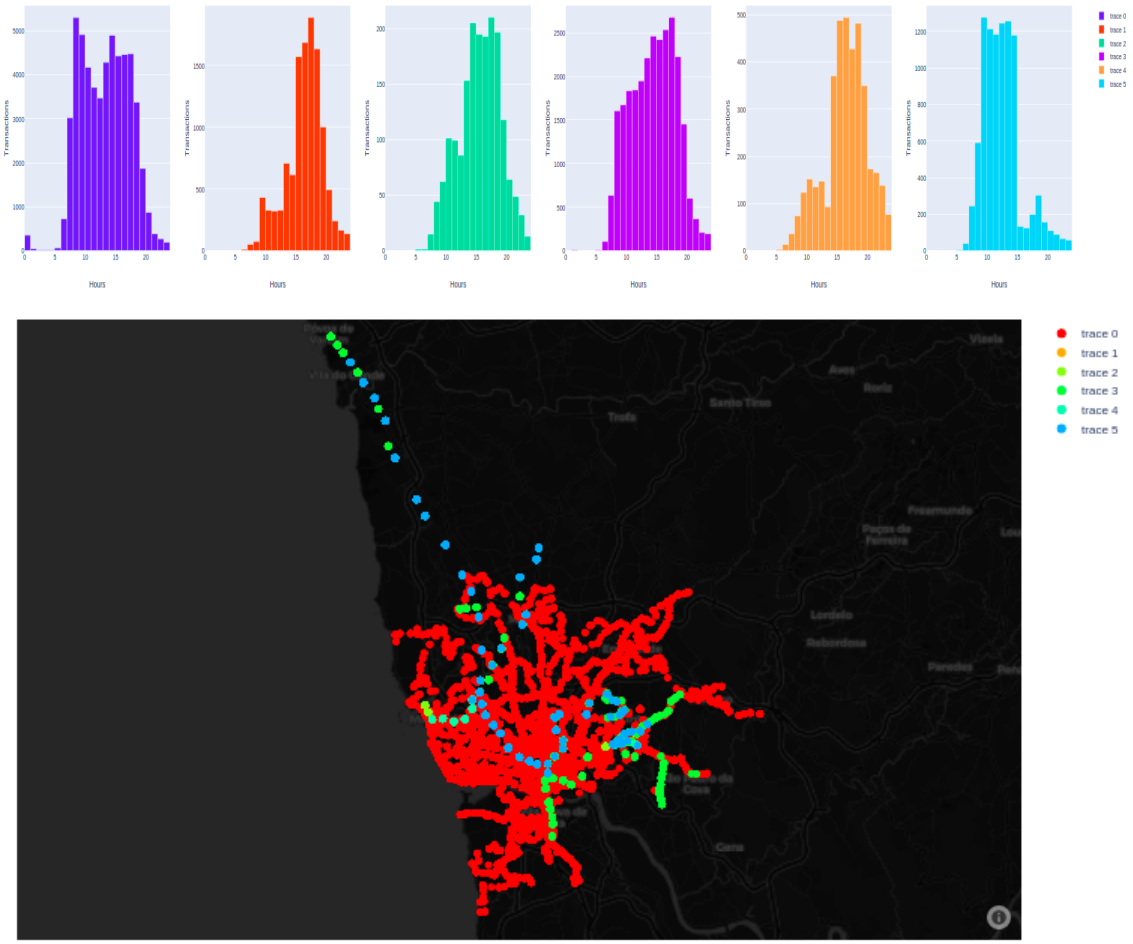


Figure 4.26: K-Means Result

4.6.3 Trip-Chain

Since many studies related to frequent passengers data in transport try to make a trip-chain model, we decided to verify in the non-frequent passenger's data if trip-chain happens. Our trip-chain approach is trip-chain of zones instead of stations. We group stations into zones mainly because many stations are from the same zone, normally the difference between those stations is that they belong to different transport lines.

To group all the transport network stations into zones we used the DBSCAN with parameters set to 0.002 for "eps" and 1 for "min_samples". We use the same value for "eps" that was used in the DBSCAN approach in spacial patterns and a value of 1 to "min_samples" because we don't want any restriction to create zones. We want the zones to depend only by distance and nothing more. This resulted in 4074 stations being clustered into 807 zones. The next image shows these zones.

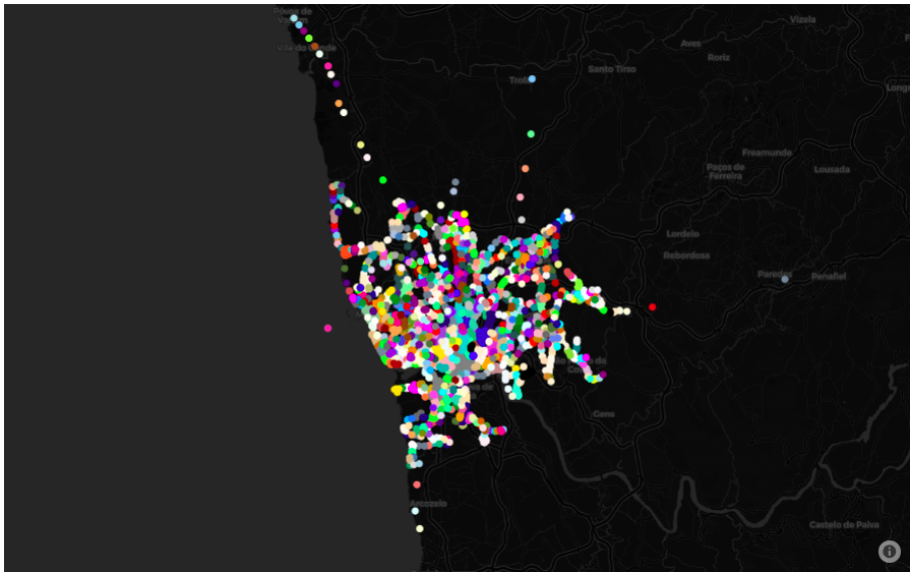


Figure 4.27: Trip-Chain Zones

Afterwards, we aggregate all the transactions that each id does in a single day. So we first get a list of all the unique ids of that day, and then gather all the transactions done by each one of those ids. In the end, we have something similar to the next Figure.

1348116055467652	[1340, 9]	[97, 13]	[35, 14]	[1569, 14]	[1340, 17]	[2781, 19]	None	None	None
1348781775398528	[1340, 9]	[97, 13]	[35, 14]	[1569, 14]	[1340, 17]	[2781, 19]	None	None	None
1350796115060352	[1236, 14]	[1798, 15]	[622, 17]	[58, 17]	None	None	None	None	None
1352682545227392	[327, 9]	[30, 9]	[1641, 11]	[1804, 11]	[3619, 12]	[1775, 12]	[1, 13]	None	None
1353695217985152	[2206, 10]	[1319, 11]	[1458, 12]	[1397, 13]	[1562, 13]	[2687, 16]	[2822, 16]	None	None
1354335168112256	[2269, 6]	[2826, 7]	[2820, 7]	[2554, 8]	[1427, 17]	[3522, 18]	None	None	None
1354616353006724	[3828, 10]	[1757, 11]	[3259, 12]	[99, 13]	[105, 14]	[99, 15]	None	None	None
1355786867058308	[2140, 6]	[29, 7]	[167, 8]	[2781, 9]	[2553, 9]	[3909, 12]	[58, 12]	[167, 14]	[21, 15]
1356252735546496	[3828, 10]	[1757, 11]	[3259, 12]	[99, 13]	[105, 14]	[99, 15]	None	None	None
1356370982610560	[2206, 10]	[1319, 11]	[1453, 11]	[1458, 12]	[1397, 13]	[1562, 13]	None	None	None
1356452586989184	[132, 8]	[2383, 8]	None	None	None	None	None	None	None
1358398207174272	[254, 15]	[333, 18]	[235, 18]	[254, 21]	[1798, 22]	[29, 22]	[29, 22]	[35, 22]	[35, 22]
1359068222072452	[2072, 7]	[1775, 8]	[1785, 16]	[1685, 16]	None	None	None	None	None
1359888560825984	[254, 10]	[2688, 11]	[1209, 13]	None	None	None	None	None	None
1361616077203072	[2163, 11]	[211, 14]	[205, 15]	None	None	None	None	None	None
1363384664204928	[462, 13]	[1746, 14]	[3055, 16]	[1796, 16]	None	None	None	None	None
1365373234062976	[199, 10]	[3146, 11]	[3790, 15]	[35, 15]	[97, 18]	[35, 18]	[21, 19]	[19, 19]	[181, 20]
1365635227068036	[360, 8]	[492, 8]	[2828, 9]	[1294, 9]	[83, 9]	[35, 10]	[35, 10]	[486, 14]	None
1366459297170304	[825, 13]	[2140, 14]	[1798, 15]	[2781, 16]	[1024, 16]	[1550, 17]	[2140, 18]	None	None
1367378983790208	[1067, 8]	[1549, 8]	[1592, 8]	[1095, 10]	[2951, 21]	[1108, 22]	None	None	None
1377231638767232	[235, 9]	[1796, 11]	[235, 14]	[1528, 15]	None	None	None	None	None
1377373372688000	[276, 8]	[274, 8]	[2383, 8]	[1745, 18]	None	None	None	None	None
1377426790098048	[1841, 11]	[1426, 14]	[3047, 14]	[1096, 14]	[1252, 17]	None	None	None	None
1377985135846528	[1841, 11]	[1832, 11]	[1426, 14]	[3047, 14]	[1096, 14]	[1252, 17]	None	None	None
1384998951658624	[99, 10]	[97, 10]	[183, 11]	[99, 12]	[3739, 15]	[3259, 17]	[1048, 18]	None	None
1387307632043648	[167, 7]	[3095, 8]	[3199, 12]	[1952, 13]	[35, 14]	[303, 16]	[3095, 16]	None	None
1387477147788416	[99, 10]	[97, 10]	[183, 11]	[99, 12]	[3739, 15]	[3259, 17]	[1048, 18]	None	None
1392806129706624	[1775, 9]	[235, 11]	[294, 12]	[108, 12]	[97, 15]	[83, 16]	[35, 17]	None	None
1396669300933248	[75, 11]	[19, 11]	[199, 12]	[199, 12]	[19, 13]	[75, 14]	[29, 15]	[35, 15]	[1750, 18]
1397231941649024	[75, 11]	[19, 11]	[199, 12]	[199, 12]	[19, 13]	[75, 14]	[29, 15]	[1750, 18]	[99, 21]

Figure 4.28: List with each ticket id transactions from February 4

In Figure 4.28, we can see all the transactions that each id has done. The first column is the id of the ticket, and the following columns are composed of the transactions that each id has done. The transaction information is composed by an array with two values in the following format ["Station id", "Hour"]. The station id the id of the station where the transaction occurred and the hour is the hour when the transaction occurred.

Afterwards, we selected from the list represented in Figure 4.28 the pairs of transactions, from each id(passenger) that occurred within 1 hour, for example from 1 pm to 2 pm. This value can be changed if we want. In this case, we decided to leave it at 1 hour, mainly, because we are trying to get the trips that happen one after other in the transport network.

Those pairs of transactions that happened within 1 hour are then saved in a list where the first column is the first station of the pair, and the second column is the second value of the pair, as Figure 4.29a shows.

Now we have the pairs of transactions that are trip chained. To understand better these trip-chains we will pass the station id value to their correspondent zone value, that we determined in the beginning with the DBSCAN as the Figure 4.29b shows. This step consists only in changing the station value to their zone value. The following figure represents the values of the trip-chained pairs list with the station values and the same list with zones values instead.

2051	2062
68	1579
1087	31
1073	109
1678	1144
199	35
177	19
8	3095
2769	1746
1746	1750
92	234
4	14
1121	1247
2122	317
317	2885
1248	1785
35	35
6	33
33	109
1982	2704
2704	105
85	3441
3441	2705
31	35
29	35
35	35
111	35
93	35
1311	1205
3238	1206

(a) List of Trip Chain Pairs with stations ids

493	530
66	39
358	39
350	42
2	0
12	42
108	32
18	42
453	42
42	42
80	132
22	29
82	82
15	89
89	42
79	42
22	41
41	42
499	673
673	88
75	88
88	673
39	42
39	42
42	42
81	42
82	82
70	82
672	485
79	42

(b) List of Trip Chain Pairs with Zones ids

Figure 4.29: Trip-Chain Pairs list before(Stations) and after(Zones)

In the end, we count how many times each unique pair of zones appears on the list. This way we can understand which pairs of zones are used the most in trip-Chain journeys.

0	1	counter
0	0	1
0	2	1
0	7	1
0	19	1
0	20	2
0	22	4
0	37	1
0	39	9
0	42	11
0	51	1
0	75	2
0	81	1
0	347	1
0	349	1
0	350	2
0	356	2
0	358	1
0	407	3
0	442	3
1	0	1
1	32	1
1	34	6
1	37	25
1	39	40
1	42	13
1	254	1
1	296	1
1	355	2
1	358	5
1	374	2

Figure 4.30: Portion of a list of Trip Chain pairs with the times that each pair appears

Chapter 5

Finding usage patterns for unfrequent users in public transports

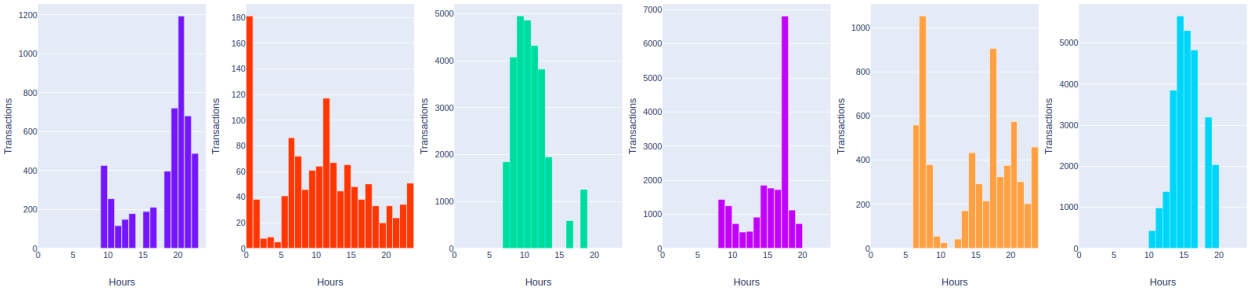
Throughout this chapter we will discuss the results of each experiment performed in Temporal Patterns, in Spatial Patterns and in Spatio-Temporal Patterns.

5.1 Temporal Results

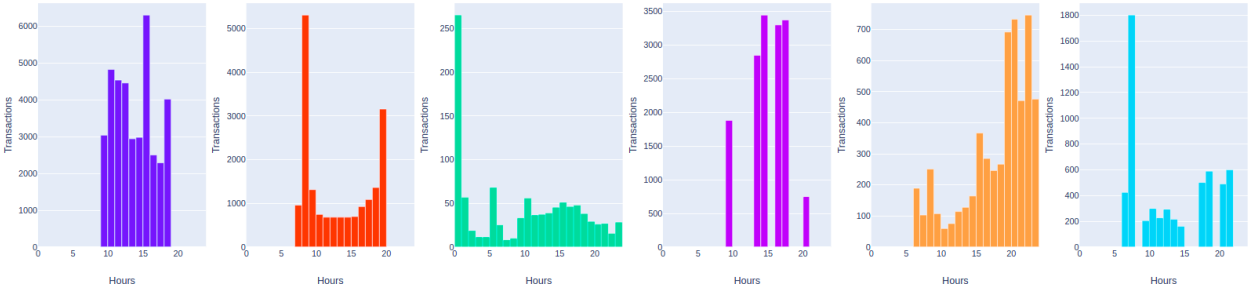
In this section, we will show the results of the approaches represented in section 4.4. These approaches were the comparison of the temporal patterns in days and weeks, and also the flow of new passengers on each day in the network.

5.1.1 Daily Data

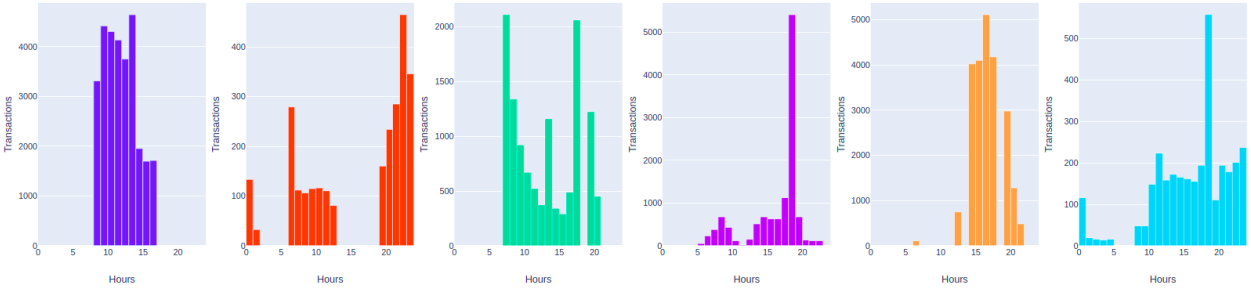
For the analysis of the daily patterns, we will show the results of each day of 4 different weeks. The first week is from 7 to 11 of January, the second week is from 8 to 12 April, the third is from 8 to 12 July and the last one is from 4 to 8 November. These weeks represent each season of the year, and we will compare the results analyse each one individually.



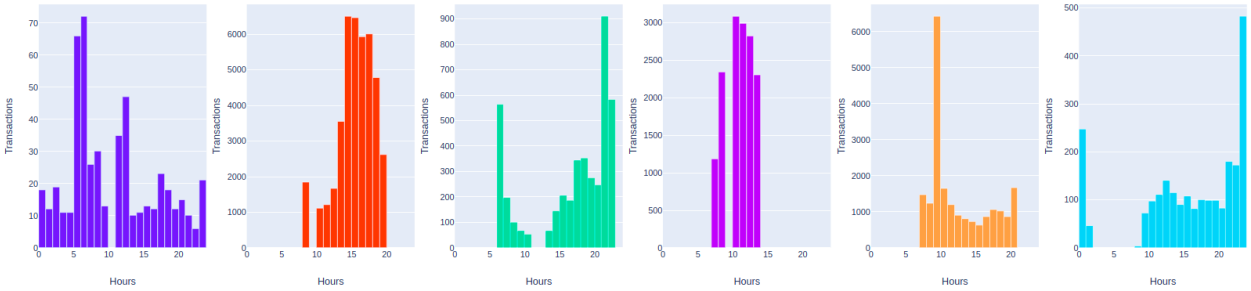
(a) January 7, 2013, Temporal Patterns



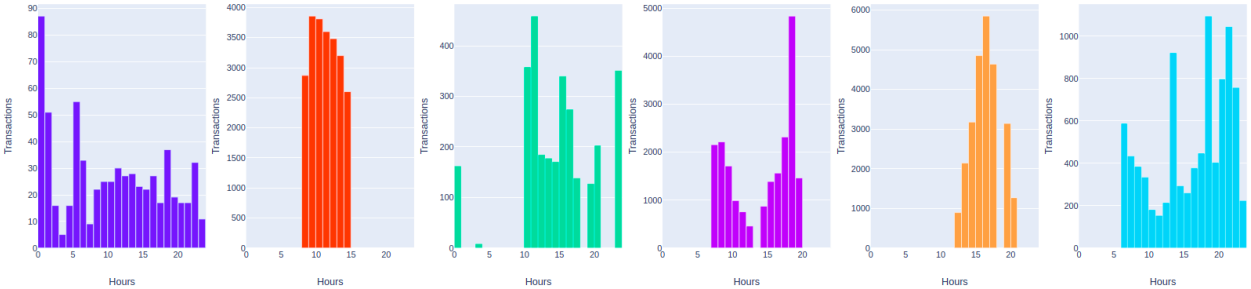
(b) January 8, 2013, Temporal Patterns



(c) January 9, 2013, Temporal Patterns



(d) January 10, 2013, Temporal Patterns



(e) January 11, 2013, Temporal Patterns

Figure 5.1: January 7 to 11, 2013, - Temporal Patterns

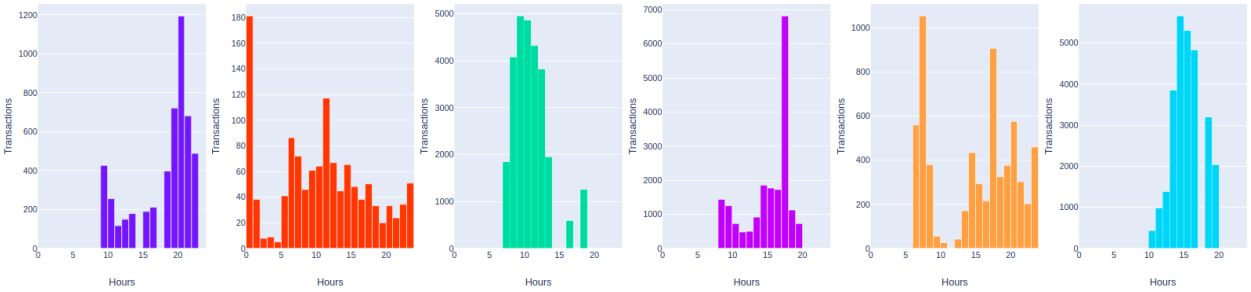
In these results, we can see that in each day 4 clusters are similar throughout the weekdays and 2 clusters that tend to have different results. First, let's talk about the 4 similar clusters that happen in each day. Those 4 clusters are composed of four different types of utilization. The first type of utilization is the morning usage of the network, the second type is the afternoon usage of the network, the third type is the morning and afternoon usage with a drop on the midday and the fourth type is the dawn usage of the network. These types can be seen on these week days. This next table will help to represent the results.

Day	Morning Type	Afternoon Type	Morning and Afternoon Type	Dawn Type
7	Green	Blue	Yellow	Red
8	N/A	N/A	Yellow	Green
9	Purple	Yellow	Green	N/A
10	Pink	Red	Green	Purple
11	Red	Yellow	Pink and Blue	Purple

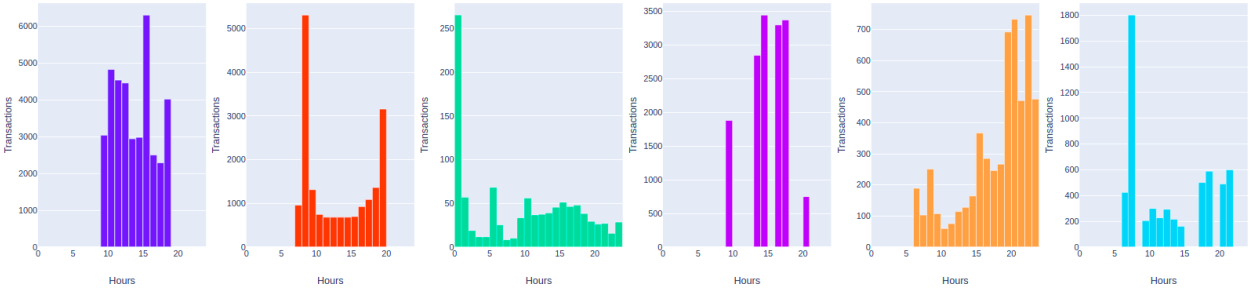
Table 5.1: Result of type of usage of January 7 to 11, 2013

These three patterns are very similar in each day, and we say very similarly because they can differ a little bit depending on the day, for example, Tuesday, 8 January. The morning and afternoon patterns seem to not exist, but in the rest of the week, they exist. This can be representative of a drop of utilization on that day or another pattern that we were not able to identify. For the other, they seem to be different to each day and hard to predict.

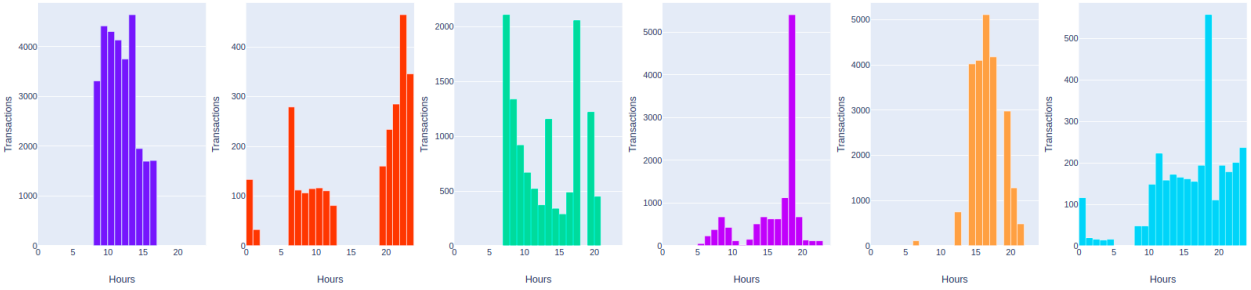
Figures5.2 to 5.4 present the results obtained for the different seasons of the year. Those seasons are Spring, Summer and Autumn since winter season is represented by the January week. So, for each season we selected a week. Those weeks were April 8 to 12 for spring, July 8 to 12 for Summer and November 4 to 8 for winter.



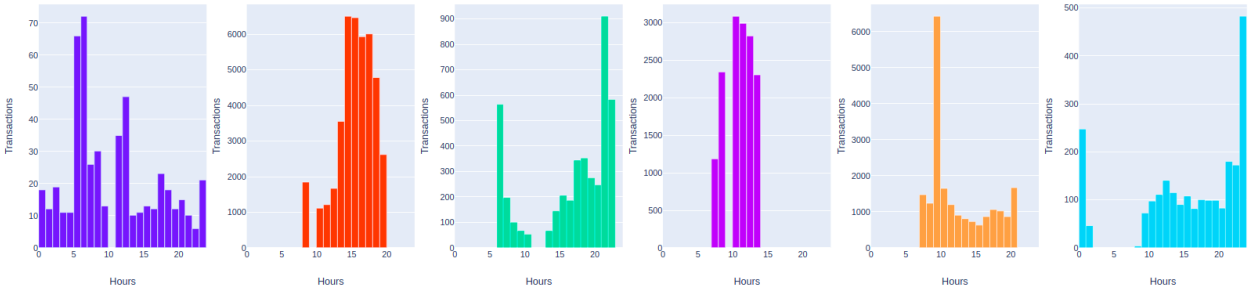
(a) April 8, 2013, Temporal Patterns



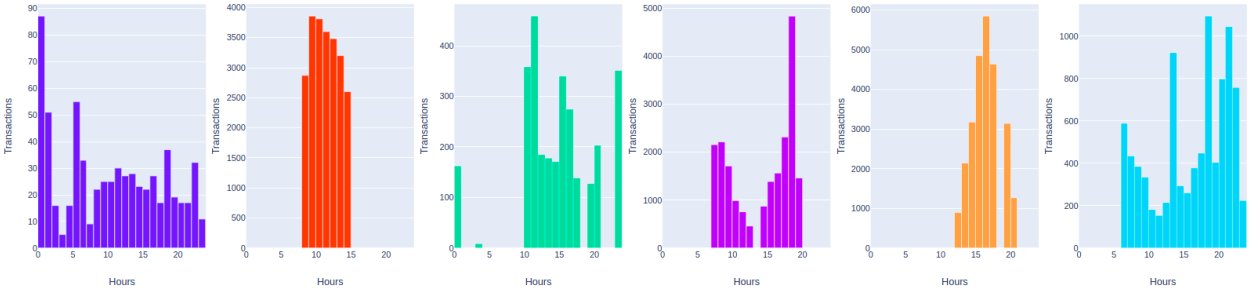
(b) April 9, 2013, Temporal Patterns



(c) April 10, 2013, Temporal Patterns

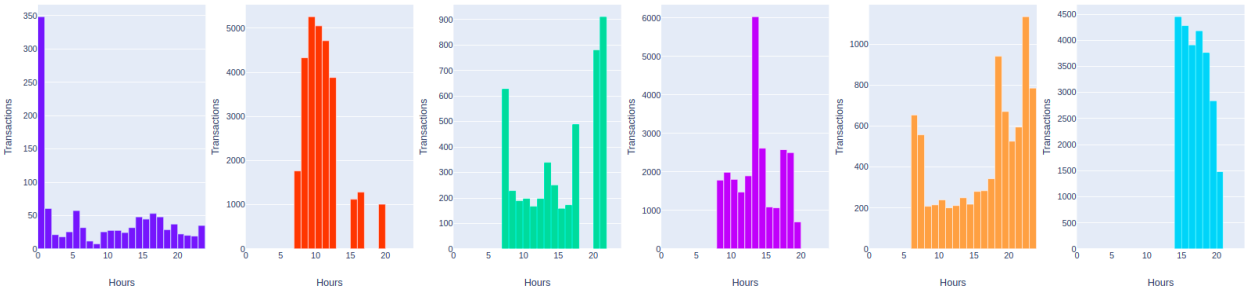


(d) April 11, 2013, Temporal Patterns

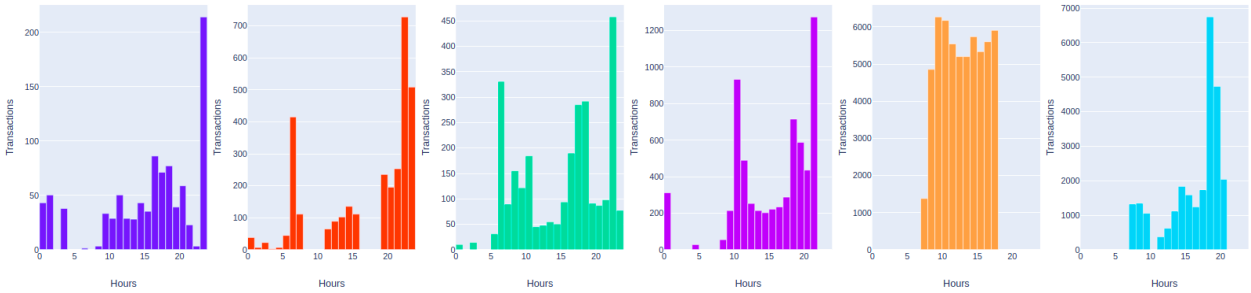


(e) April 12, 2013, Temporal Patterns

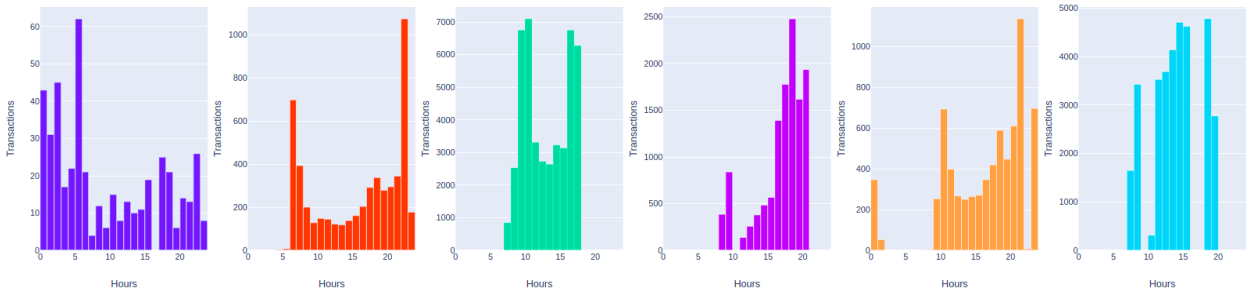
Figure 5.2: April 8 to 12, 2013, - Temporal Patterns



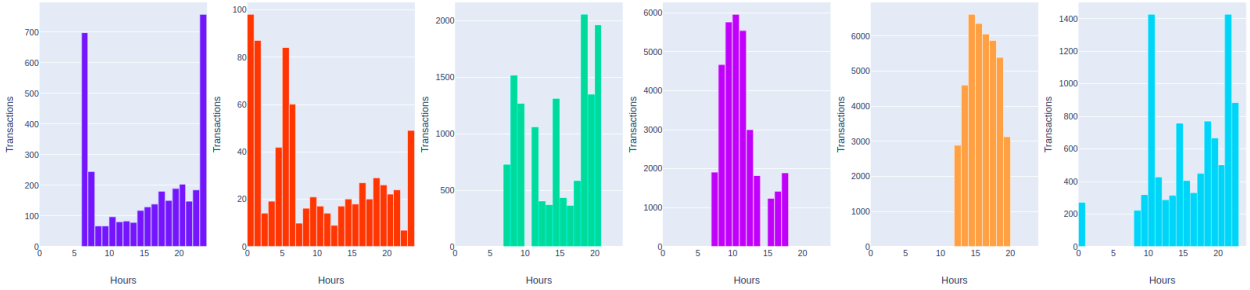
(a) July 8, 2013, Temporal Patterns



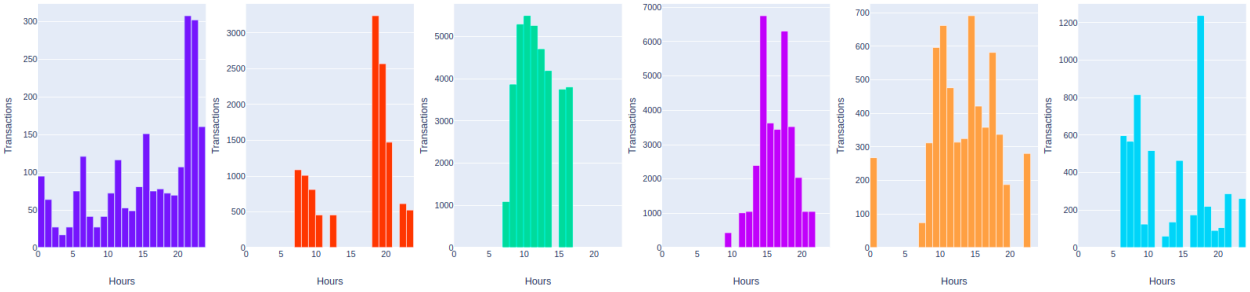
(b) July 9, 2013, Temporal Patterns



(c) July 10, 2013, Temporal Patterns

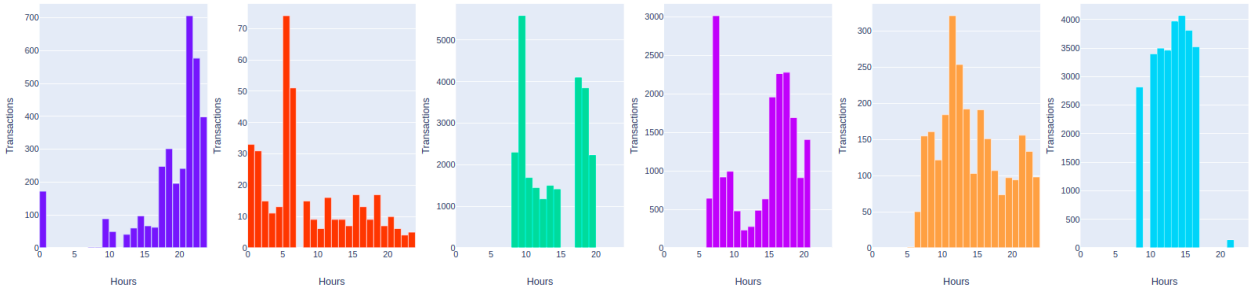


(d) July 11, 2013, Temporal Patterns

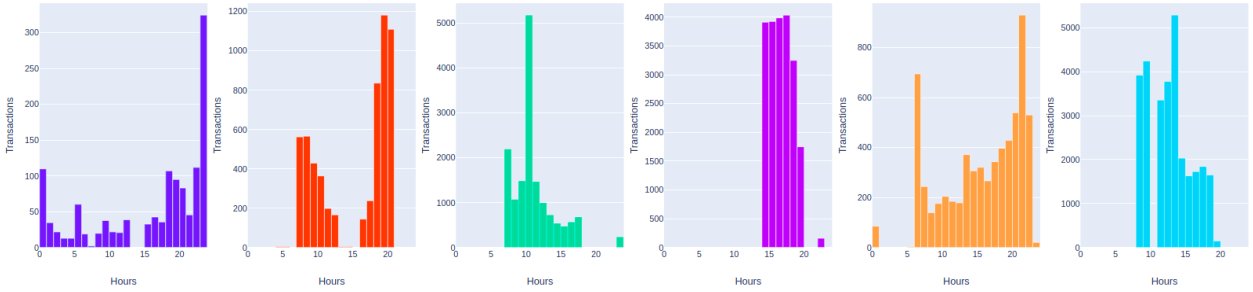


(e) July 12, 2013, Temporal Patterns

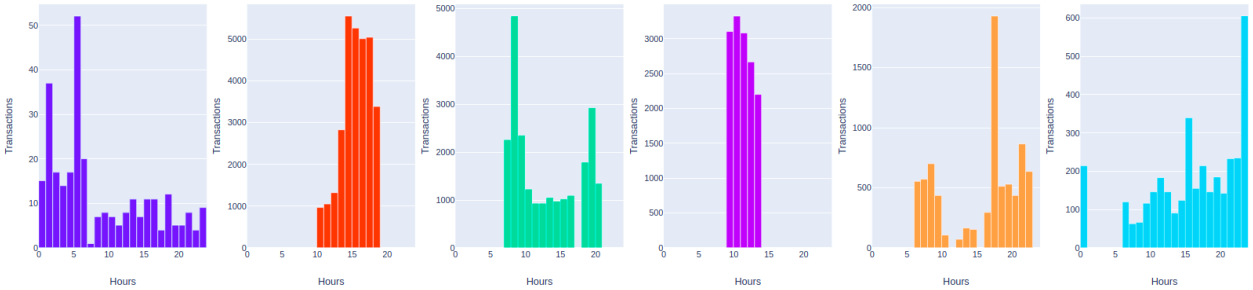
Figure 5.3: July 8 to 12, 2013, - Temporal Patterns



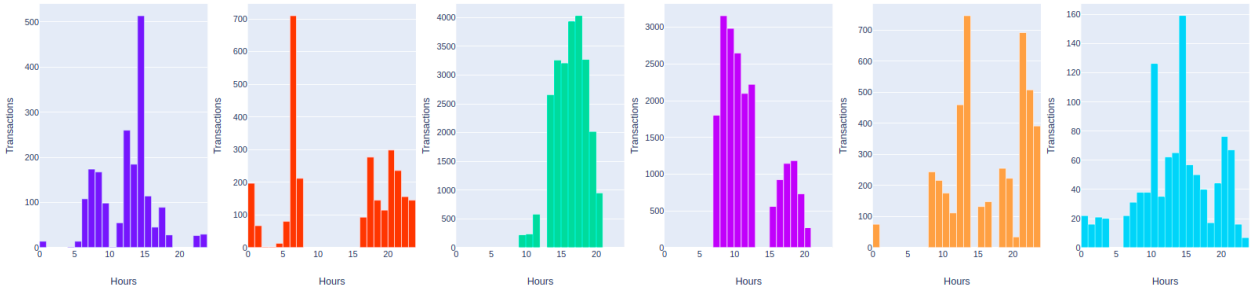
(a) November 4, 2013, Temporal Patterns



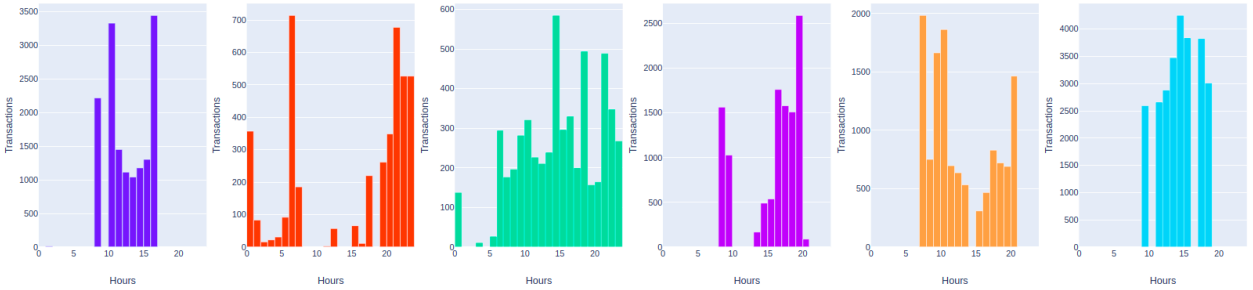
(b) November 5, 2013, Temporal Patterns



(c) November 6, 2013, Temporal Patterns



(d) November 7, 2013, Temporal Patterns



(e) November 8, 2013, Temporal Patterns

Figure 5.4: November 4 to 8, 2013, - Temporal Patterns

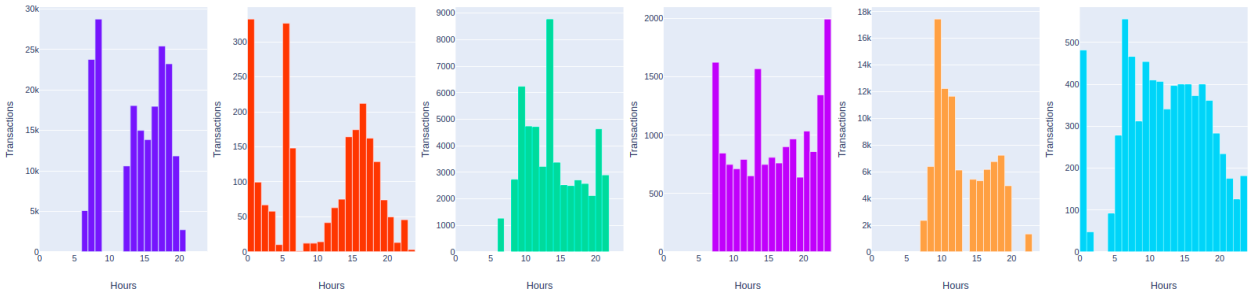
We are gonna show a table similar to Table 5.1 for these weeks that represent different seasons.

Day	Morning Type	Afternoon Type	Morning and Afternoon Type	Dawn Type
April 8	Green	Blue	Yellow	Red
April 9	N/A	N/A	Yellow And Red	Green
April 10	Purple	Yellow	Green	N/A
April 11	Pink	Red	Green	Purple
April 12	Red	Yellow	Pink	Purple
July 8	Red	Blue	Yellow and Green	Purple
July 9	N/A	N/A	Pink And Green	N/A
July 10	N/A	Pink	Green and Yellow	Purple
July 11	Pink	Yellow	Green and Blue	Red
July 12	Green	Pink	Yellow	Purple
November 4	N/A	N/A	Green and Pink	Red
November 5	Blue	Pink	Red and Yellow	Green
November 6	Pink	Red	Green and Yellow	Purple
November 7	N/A	Green	Pink and Yellow	Red
November 8	N/A	Blue	Pink and Yellow	Red

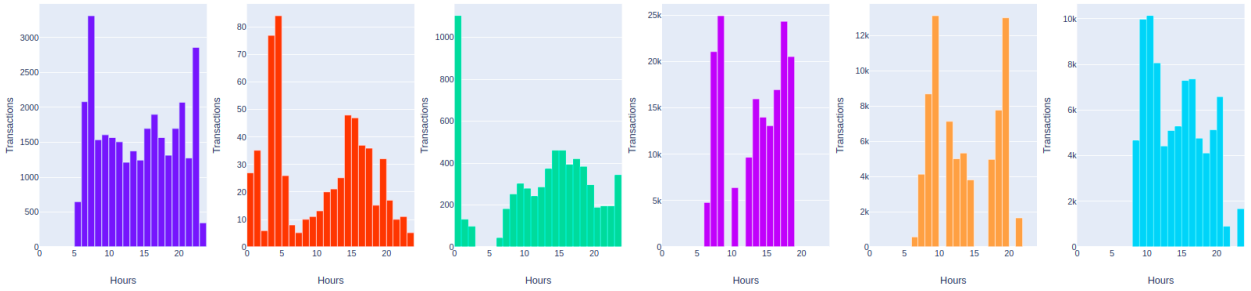
Table 5.2: Result of type of usage of different seasons representatives

As we can see in Table 5.2 there are the same patterns throughout the seasons of the year. In some particular cases, there are days where some of the patterns are harder to see, but in overall we can see in all seasons representatives the Morning, the Afternoon, the Morning and Afternoon and the dawn types of usage in the network.

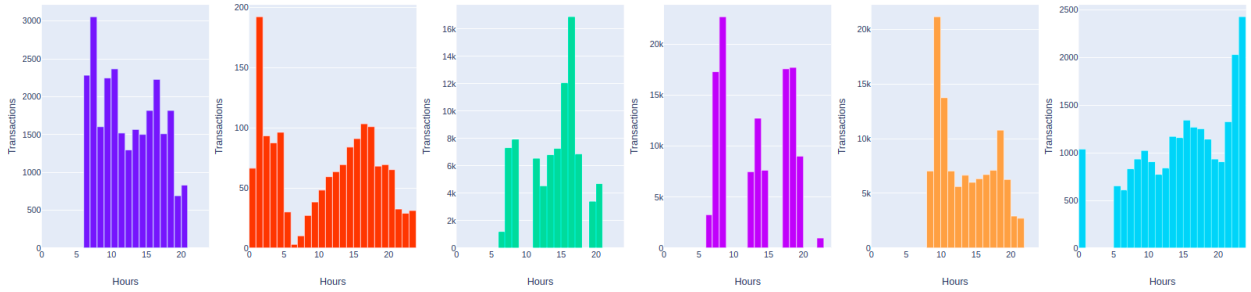
in order to test if these results are worth and correspond to real patterns, we repeated the process on frequent users data. The example that we will show here is from January 7 to January 11. The same week that we used to represent the winter season on non-Frequent passengers.



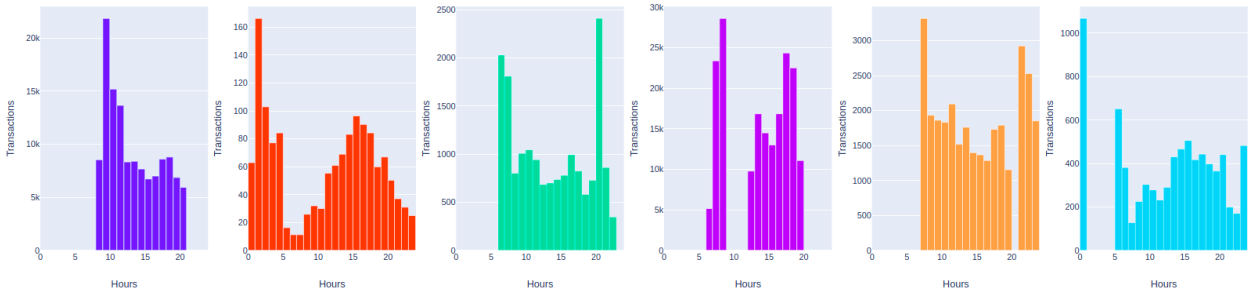
(a) January 7, 2013, Temporal Patterns of Frequent Passengers



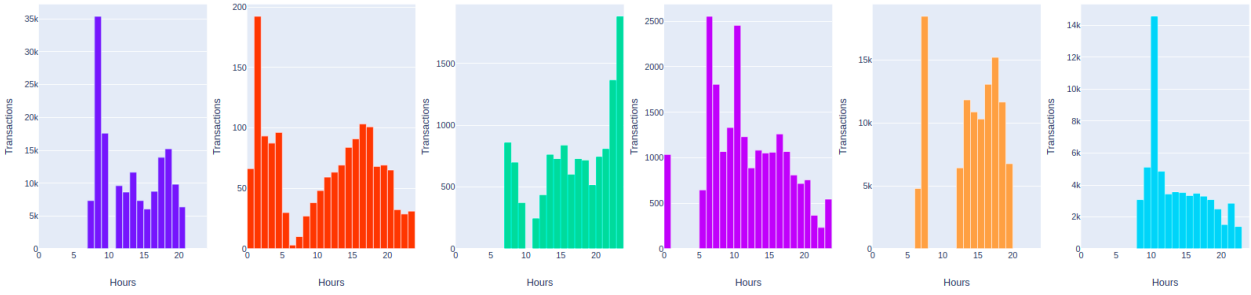
(b) January 8, 2013, Temporal Patterns of Frequent Passengers



(c) January 9, 2013, Temporal Patterns of Frequent Passengers



(d) January 10, 2013, Temporal Patterns of Frequent Passengers



(e) January 11, 2013, Temporal Patterns of Frequent Passengers

Figure 5.5: January 7 to 11, 2013, - Temporal Patterns of Frequent Passengers

The process also found some patterns on the frequent passengers data. Those patterns are the high morning usage of the network and stabilizes throughout the day, the morning and afternoon pikes of usage and the dawn usage. The next table will show the results of these types of usage.

Day	High Morning with stabilized Afternoon Type	Morning and Afternoon Type	Dawn Type
7	Yellow	Purple	Red
8	Blue and Purple	Pink and Yellow	Red
9	Yellow	Pink and Green	Red
10	Purple	Pink and Green And Yellow	Red
11	Purple and Blue and Pink	Yellow	Red

Table 5.3: Result of the type of usage of Frequent Passengers from January 7 to 11, 2013

As we can see in the graphs there are similar patterns to the non-frequent passengers. Those similar patterns are the morning and afternoon type of usage and the dawn type of usage. On the another hand, the high usage on the morning and a more stabilized in the afternoon are only common in the frequent passengers.

5.1.2 Weekly Data

Here we are going to discuss the results of temporal patterns on Week Data. For this type of data we have the results from 3 weeks of January and 3 other weeks to represent the seasons of the year. Those weeks from January are 7 to 11, 14 to 18 and 21 to 25. From the rest of the year are April 8 to 12, July 8 to 12 and November 4 to 8.

The results are interesting since they share some patterns with daily data. The results for weekly patterns are presented in Figures 5.6 to 5.11.

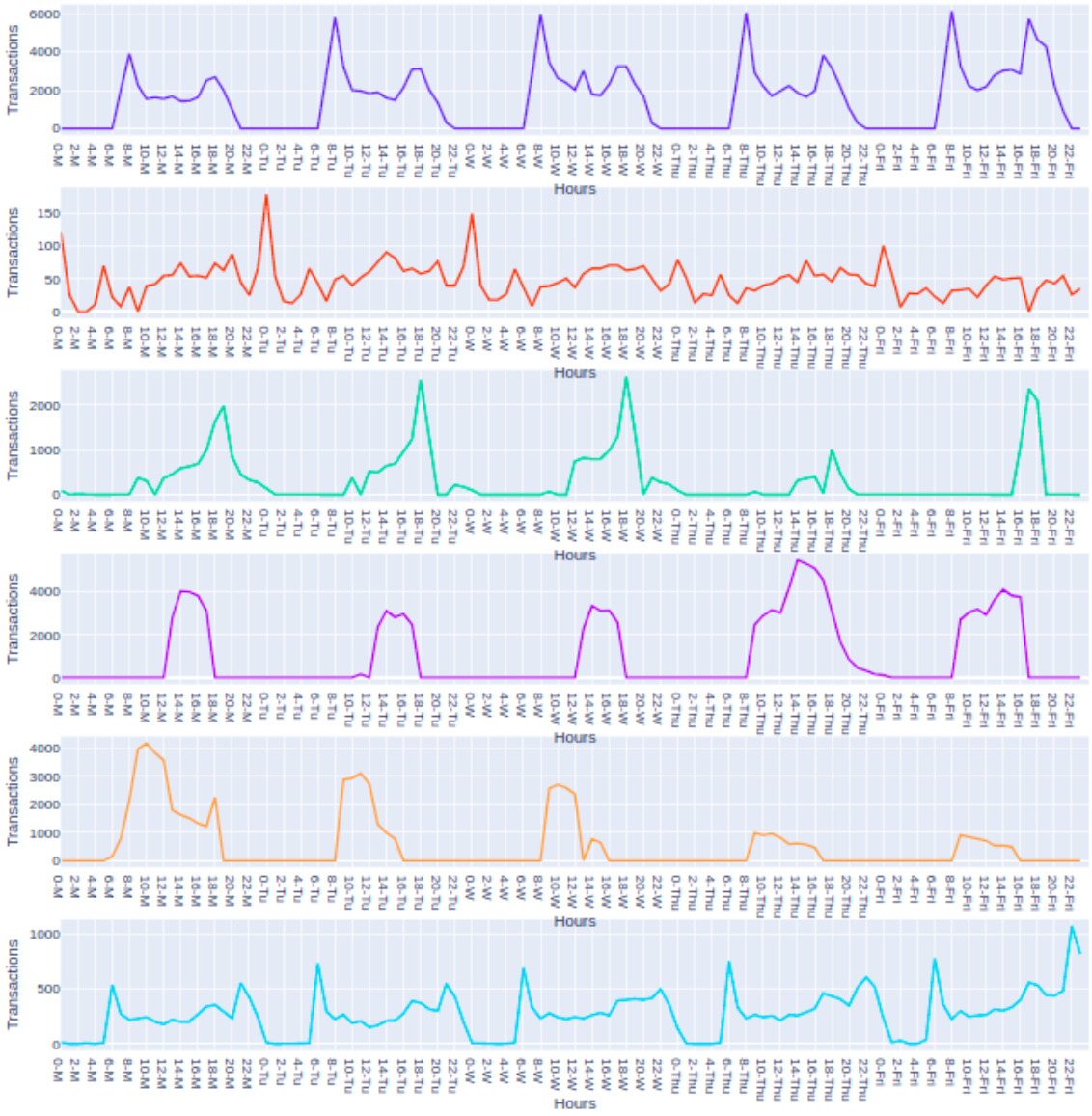


Figure 5.6: January 7 to 11, 2013 - Week Temporal Patterns

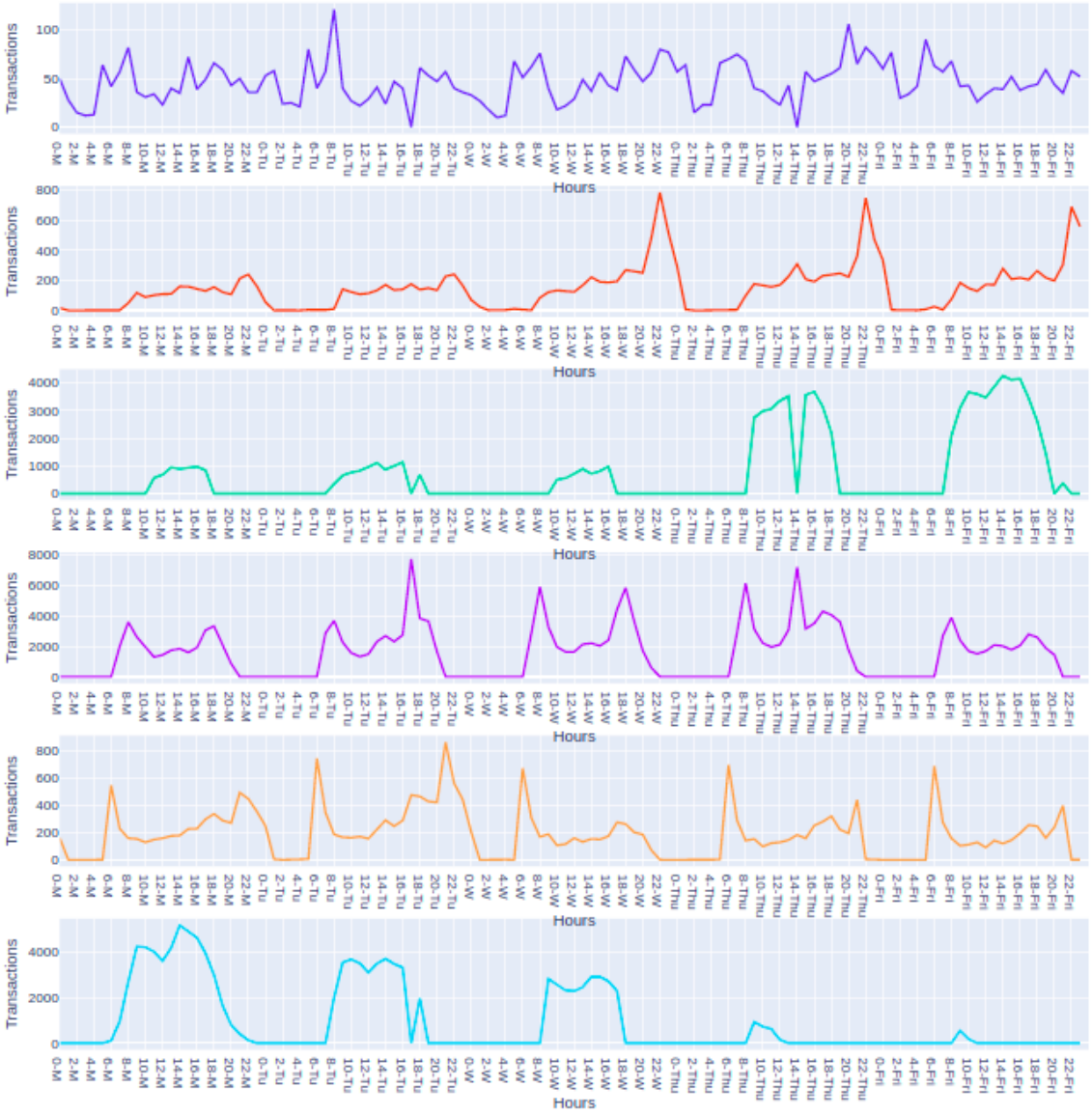


Figure 5.7: January 14 to 18, 2013 - Week Temporal Patterns

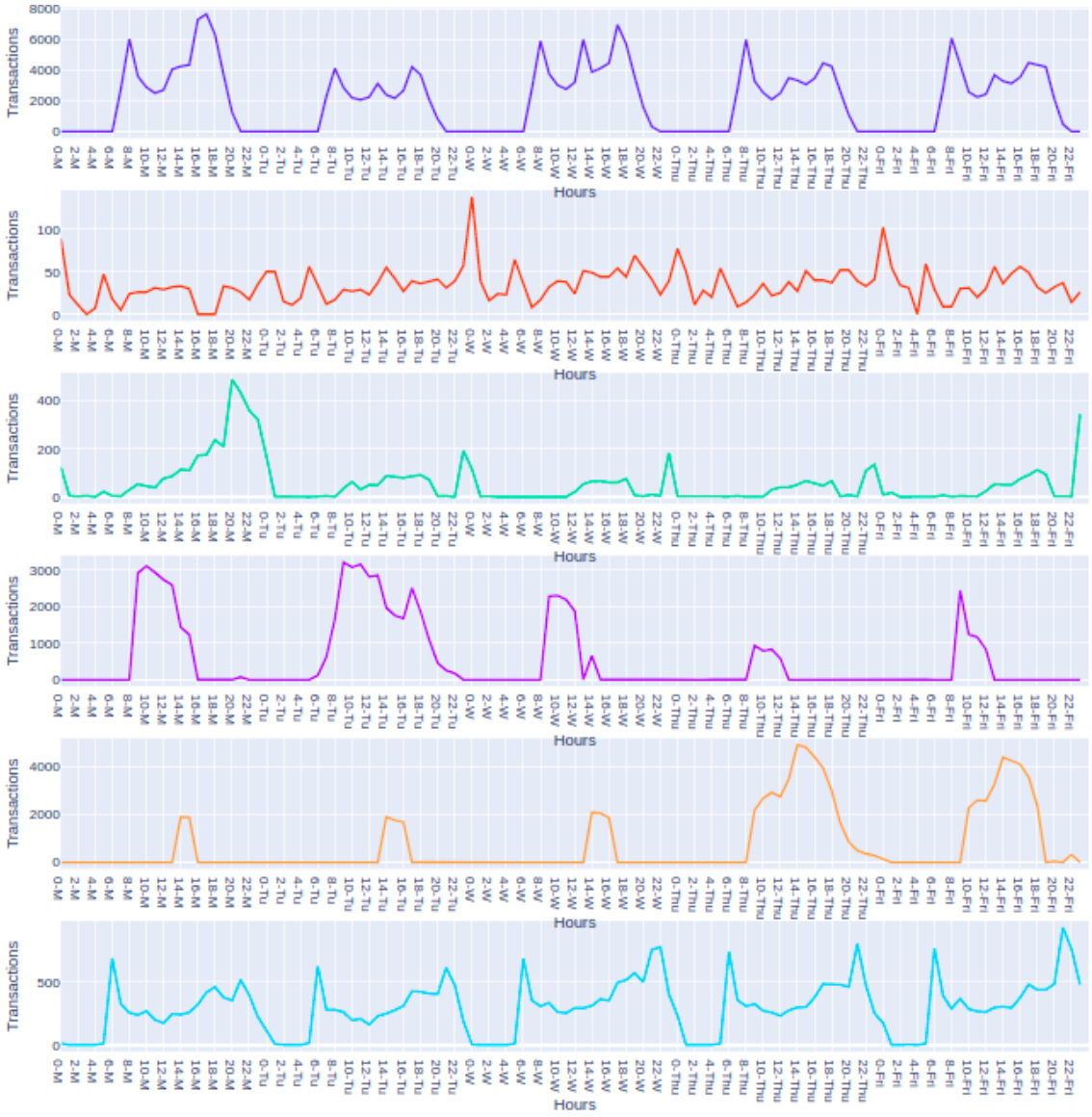


Figure 5.8: January 21 to 25, 2013 - Week Temporal Patterns

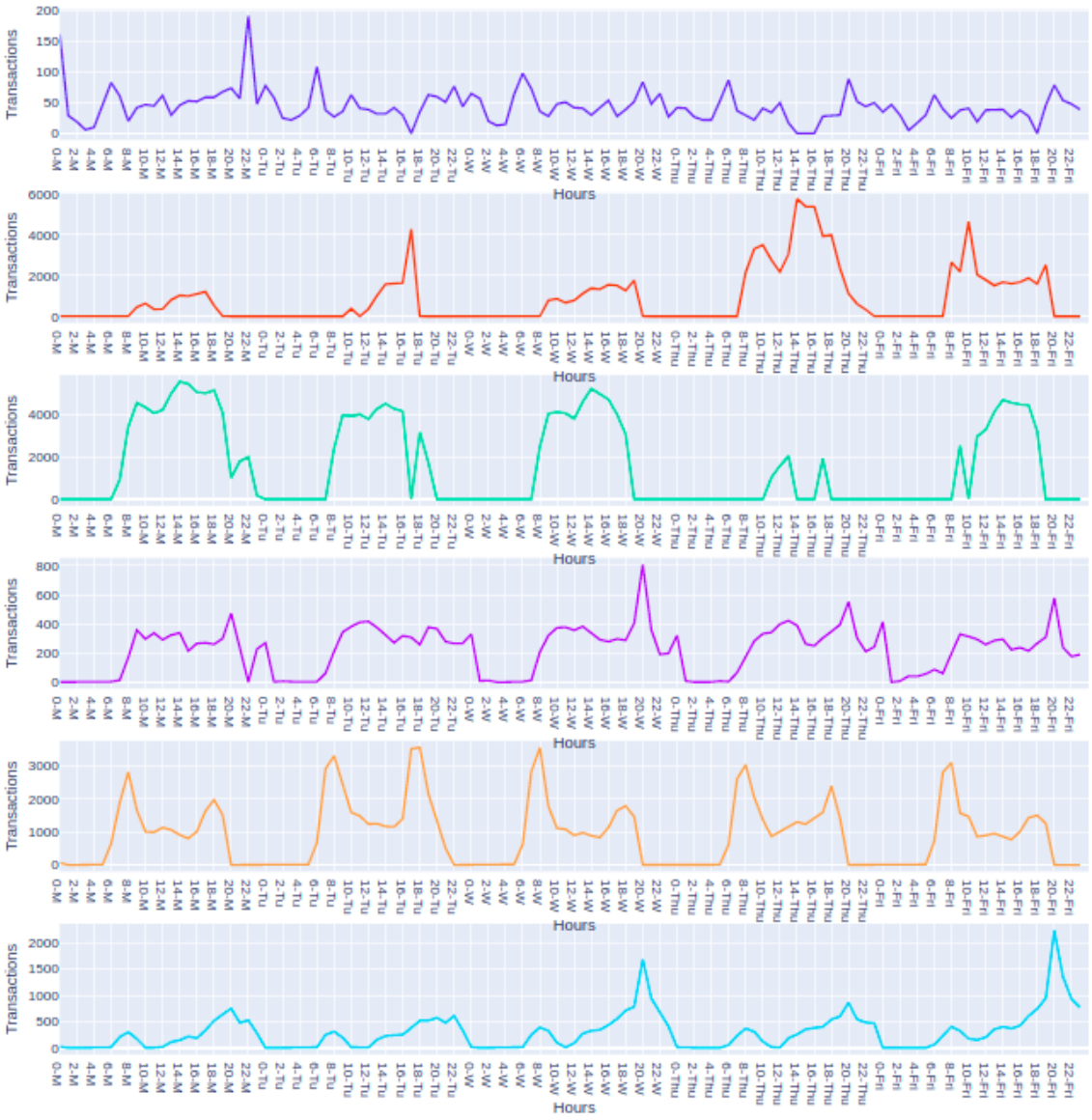


Figure 5.9: April 8 to 12, 2013 - Week Temporal Patterns

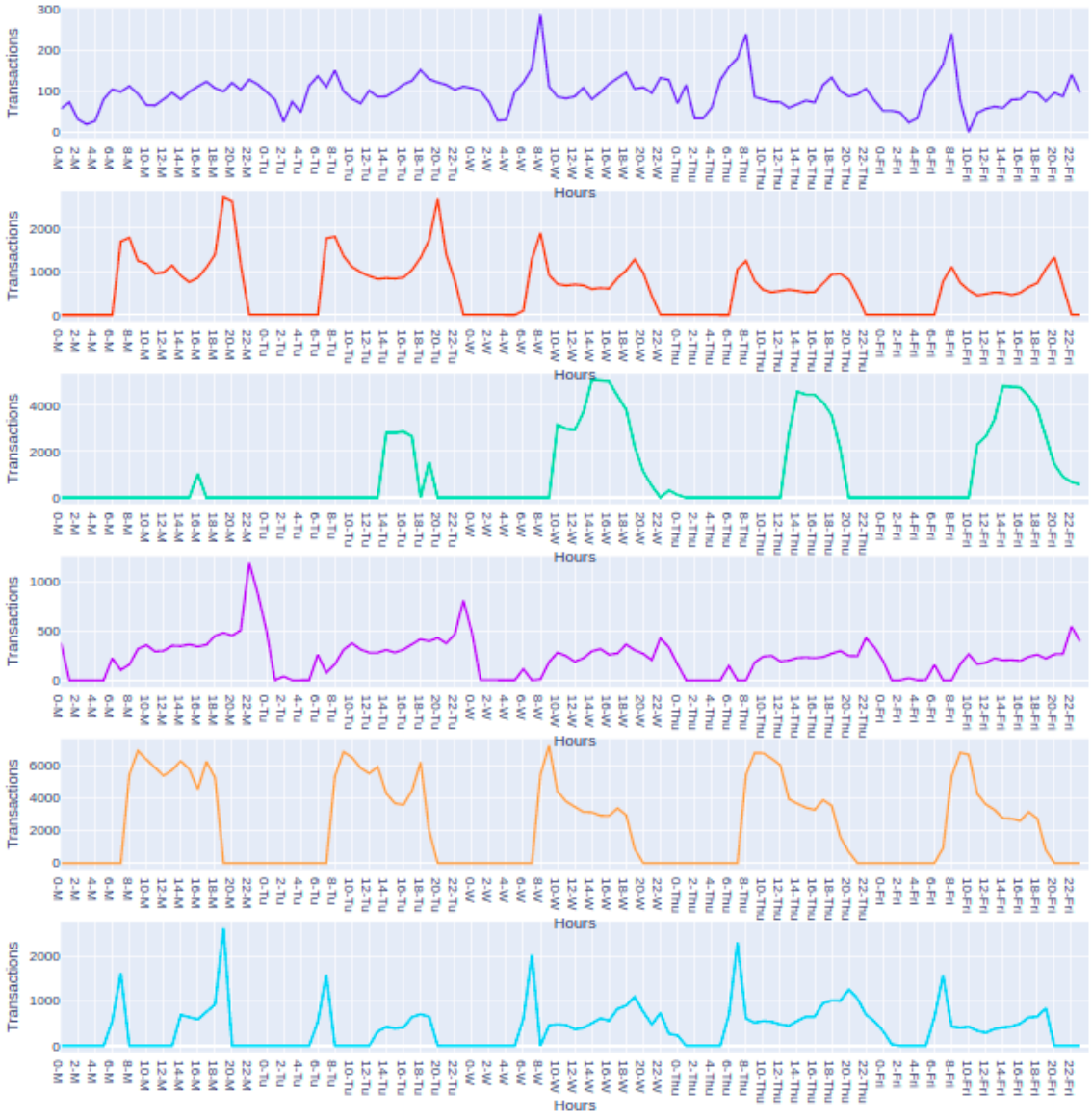


Figure 5.10: July 8 to 12, 2013 - Week Temporal Patterns

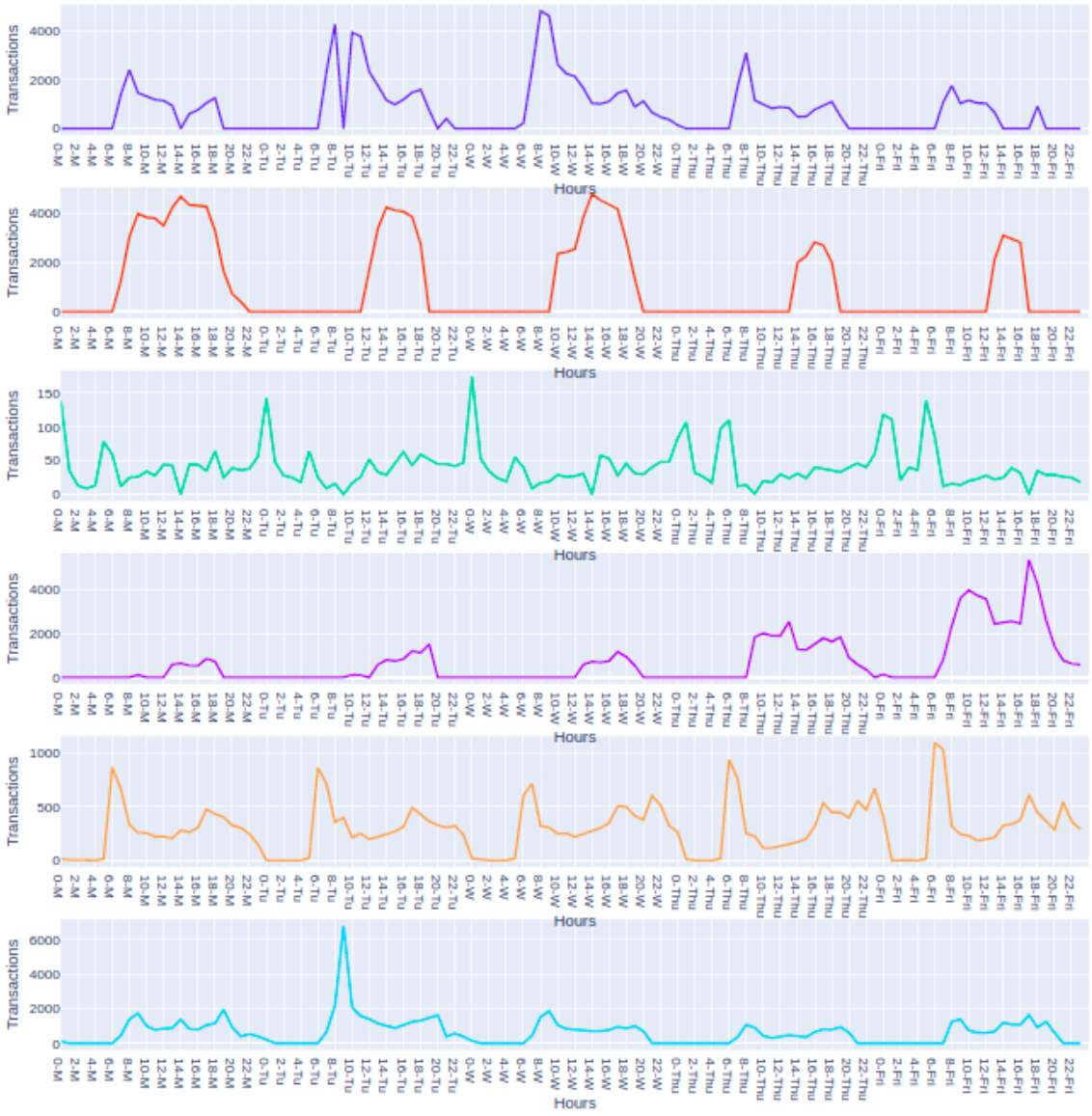


Figure 5.11: November 4 to 8, 2013 - Week Temporal Patterns

The 5.4 table show the patterns encountered in these weeks.

Week	Morning Type	Afternoon Type	Morning and Afternoon Type	Dawn Type
January 7 to 11	N/A	Pink and Green	Purple and Blue	Red
January 14 to 18	N/A	N/A	Purple and Yellow	Red
January 21 to 25	N/A	Yellow	Purple and Blue	Red
April 8 to 12	N/A	Blue	Green and Yellow and Pink	Purple
July 8 to 12	N/A	N/A	Red and Yellow and Blue	Pink
November 4 to 8	N/A	N/A	Purple and Red and Yellow	Green

Table 5.4: Result of the type of usage of Week Temporal Patterns

The results show that Morning and Afternoon types of usage are not common in week data. This could mean that those patterns are particular choices that passengers do on some days but not throughout the week. The morning and afternoon type is the most common with different types of usage inside it. We can see that some of these patterns show a peek of usage on the early morning and at the end of the afternoon. On the other hand we have some patterns that have similar usage throughout the day. These results show also some irregular clusters or clusters with patterns in certain days but not throughout the week. This last description could be a pattern, but we don't have results to ensure that.

5.1.3 New Passengers

In this section we present the results of our approach to know how the network is used by new passengers each day. The results from the first two months are a bit inflated because our pool of existent ids starts empty on the first day of January. So, until February we can see in the results some strange values since some ids are being used for the first time in our study, but not the first time to the network. So, we will show the number of new passengers and the ratio of passengers that are new, in each day for a year.

In the following figures, the first column is the day of the month, the second is the value of new passengers and the third is the ratio of new passengers. We will present the first 3 months and the following months will be on appendix.

Day	New Tickets	Day Ratio of New Tickets
1	25237	1.000000
2	52265	0.928264
3	42284	0.713016
4	33555	0.572211
5	22128	0.595223
6	11115	0.501263
7	27059	0.493498
8	22079	0.425054
9	19671	0.382110
10	19304	0.355631
11	19201	0.340407
12	10902	0.371955
13	6700	0.336650
14	16554	0.320752
15	14593	0.292550
16	12022	0.272651
17	12924	0.262001
18	11903	0.255243
19	7389	0.300537
20	4131	0.262920
21	13132	0.261573
22	10208	0.232778
23	10090	0.223621
24	12138	0.231509
25	12419	0.229323
26	9371	0.282480
27	3934	0.232026
28	13573	0.245013
29	11278	0.217332
30	11360	0.214364
31	11505	0.207608

(a) January New Passengers

1	11412	0.302401
2	9675	0.281758
3	5761	0.252875
4	12893	0.233980
5	11273	0.213649
6	10652	0.207990
7	10546	0.202609
8	11700	0.209265
9	9148	0.258360
10	5021	0.258362
11	10704	0.231683
12	5596	0.217422
13	11080	0.210275
14	10443	0.190212
15	10513	0.188024
16	7211	0.222280
17	3510	0.199557
18	9823	0.188097
19	12543	0.217338
20	8856	0.169415
21	8206	0.160842
22	8360	0.161406
23	7891	0.221043
24	4104	0.191480
25	8880	0.165258
26	8322	0.153724
27	8114	0.150156
28	8578	0.150592

(b) February New Passengers

1	11925	0.189259
2	9498	0.240249
3	4319	0.194690
4	8413	0.162197
5	6809	0.164033
6	7561	0.152483
7	6716	0.144748
8	9395	0.159632
9	6060	0.195579
10	3546	0.183398
11	7550	0.151467
12	7534	0.152288
13	7910	0.152717
14	7644	0.146269
15	8581	0.152806
16	5747	0.191835
17	4401	0.201566
18	8397	0.165906
19	8367	0.159108
20	8369	0.158636
21	8353	0.157800
22	8159	0.155555
23	6003	0.188270
24	4149	0.208744
25	7058	0.173266
26	7183	0.163102
27	8266	0.167101
28	8722	0.171255
29	5385	0.234457
30	5800	0.196384
31	3159	0.214869

(c) March New Passengers

These results show a consistent number of new passengers or at least new tickets each day. The lowest value that we got throw the year of new passengers was 970 on December 25 and that value corresponds to 16 per cent of all non-frequent passengers that use the network on that day. The results show an average of new passengers per day throughout the year of 7532 passengers. This average drops to 6498 passengers when we remove the two first months that we thing that are inflated. Now, for the average of the ratio, the value maintains almost equal with the 2 first months and without them. The average is 15.7 with all months and 15.8 without the two first months of the year. So, to sum up, the network as an average of 7532 passengers new passengers per day, that normally, corresponds to an average of 15.7 per cent of the total occasional passengers that

use the network, per day.

5.2 Spatial Results

In this section, we will show the results of the approaches represented in section 4.5. In these approaches we used DBSCAN to find the places with more affluence in the network and also the most frequently used type of ticket.

5.2.1 DBSCAN Results

The search for the most common areas in the network was done by day. This way we could see, each day, what are the areas used the most. We made a comparison of the results for a week and then a comparison of the results of different seasons of the year, by picking the same weekday on a different season.

First, we are going to compare the results of the most common areas of a week. The week in the study is January 7 to 11. The results are plotted in a Map of the city of Porto for a better visualization of the areas.

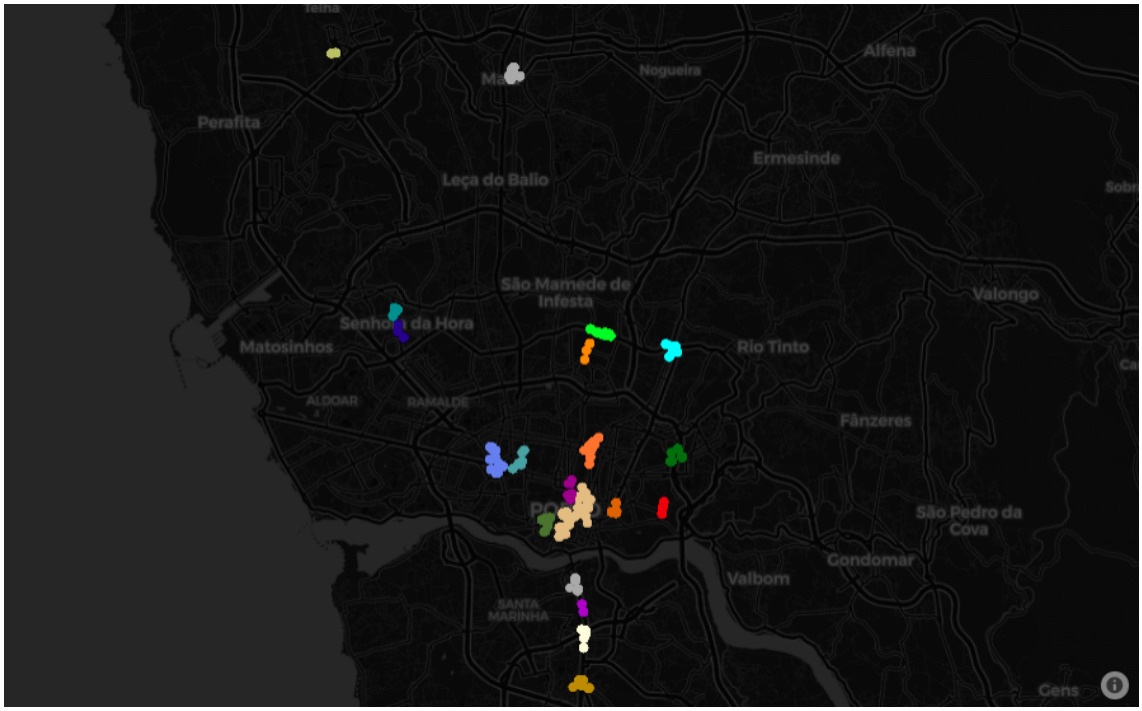


Figure 5.13: January 7 DBSCAN Results - 20 Clusters

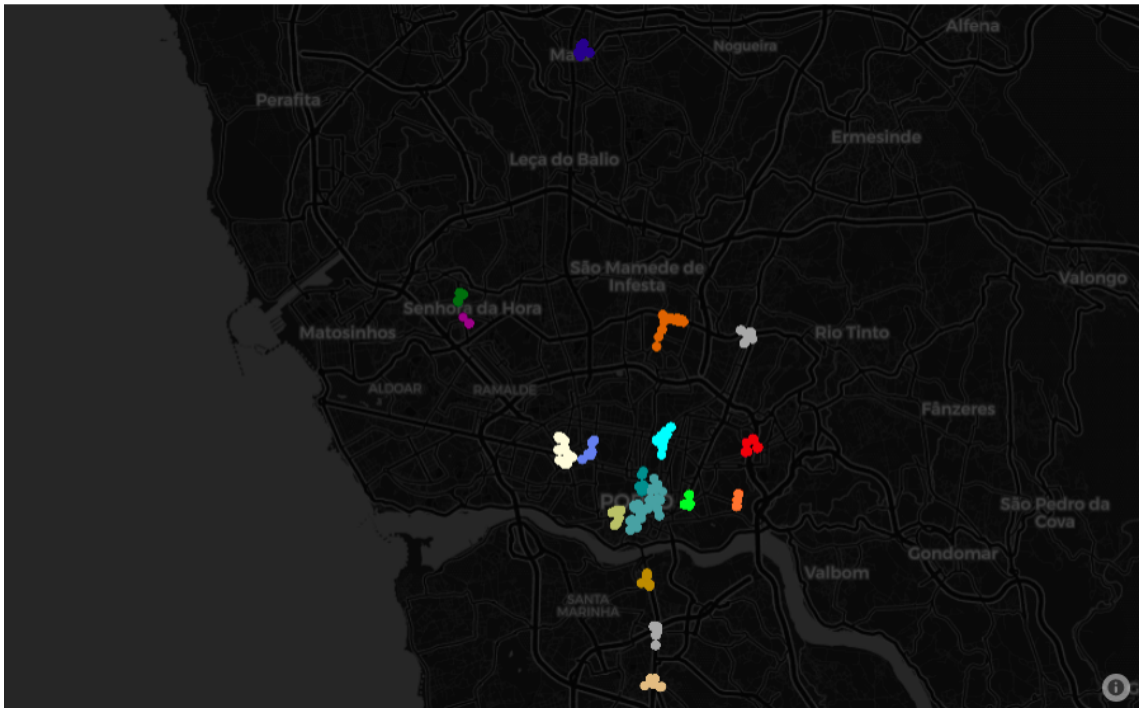


Figure 5.14: January 8 DBSCAN Results - 17 Clusters

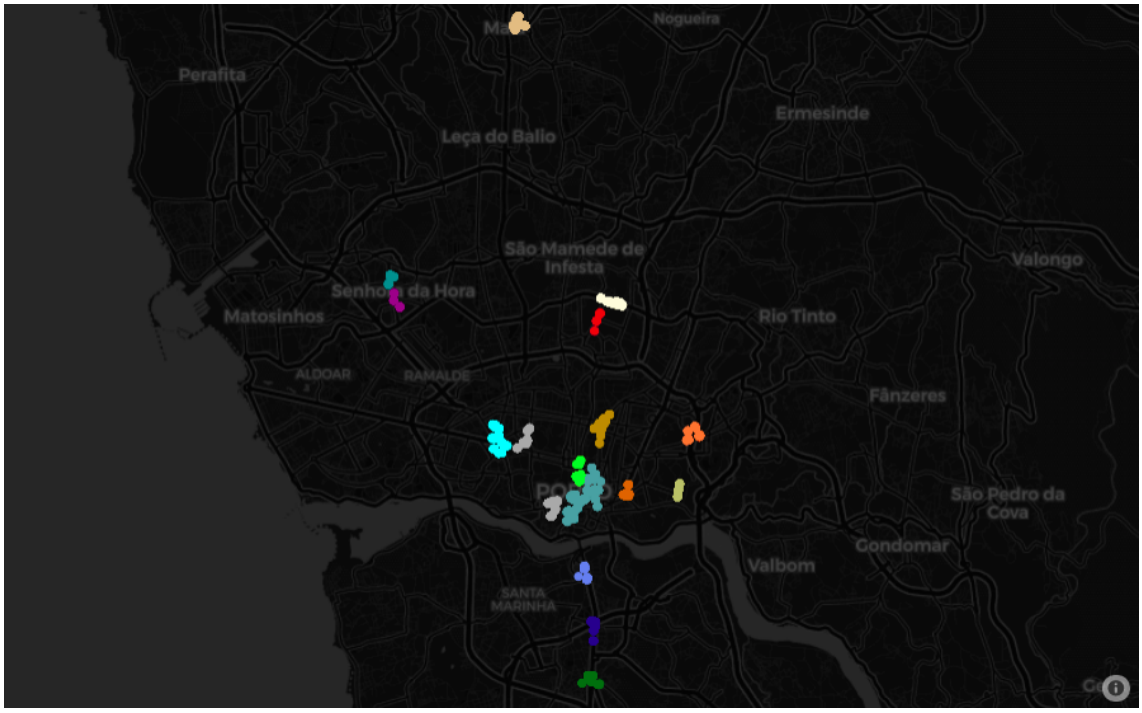


Figure 5.15: January 9 DBSCAN Results - 17 Clusters

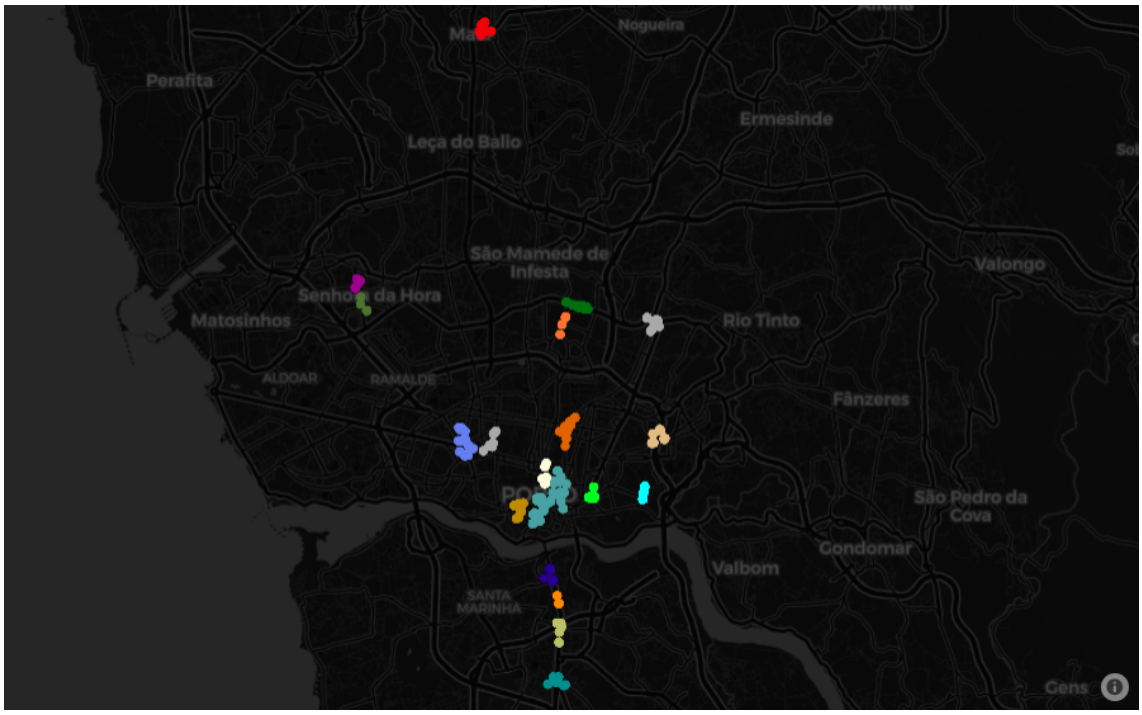


Figure 5.16: January 10 DBSCAN Results - 19 Clusters

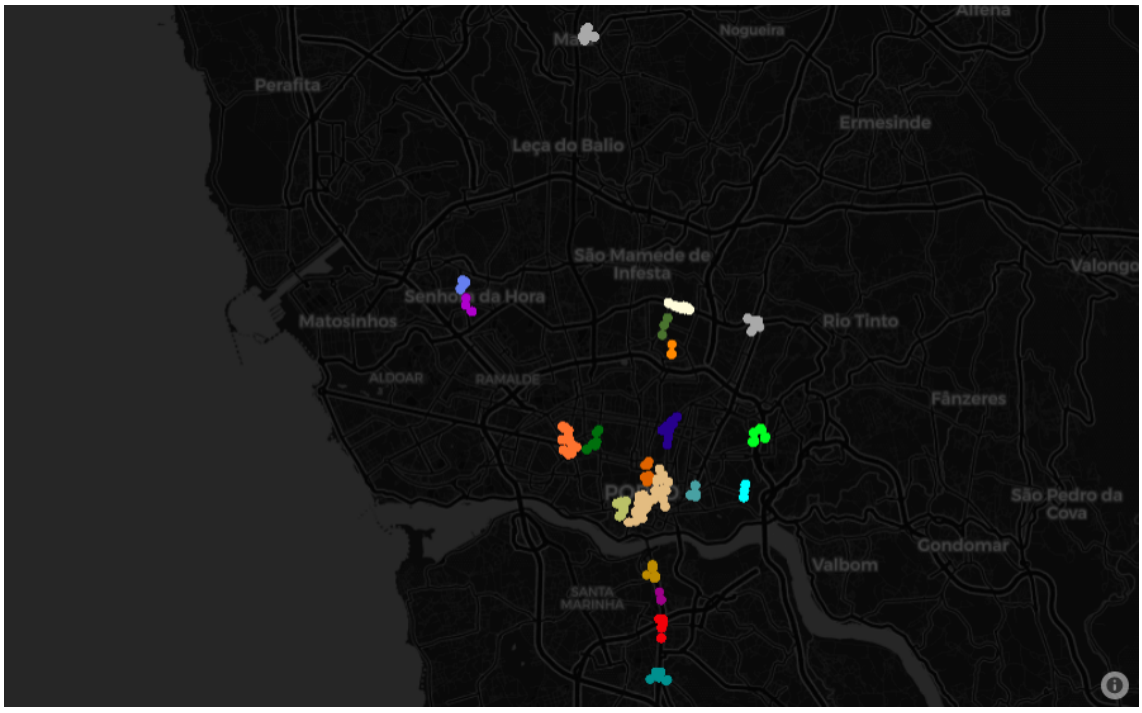


Figure 5.17: January 11 DBSCAN Results - 20 Clusters

These are the results of the January week from 7 to 11. The results show very similar common areas throughout the week. The result with fewer clusters(8 and 9 of January) has 17 and the one with the most has 20(7 and 11 of January). We will represent the areas with common names of the transport network. This next list will show those areas and represent if they are common or if they only appear sometimes.

Location	Common/Sometimes
Maia	Common
Senhora Da Hora Metro Station	Common
Norte Shopping	Common
Hospital São João	Common
IPO Metro Station and Universities	Common
Areosa	Common
Cordoaria	Common
Aliados and Bolhão	Common
Trindade	Common
Rotunda da Boavista	Common
Marquês	Common
Carolina Michaelis Metro Station	Common
Campo 24 de Agosto Metro Station	Common
Jardim do Morro	Common
General Torres	Sometimes
El Corte Inglés	Common
Santo Ovidio	Common
Estádio do Dragão	Common
Campanhã	Common
Francisco Sá Carneiro Airport	Sometimes

Table 5.5: Most Used Areas of the Network On January 7 to 11

This list shows that the most commonly used areas are the change of line or transport areas, like Rotunda da Boavista, Francisco Sá Carneiro Airport, Campo 24 de Agosto, Campanhã, Santo Ovidio, General Torres And Trindade, the shopping's areas as NorteShopping and El Corte Inglés, the touristic areas in particular Marquês, Estádio do Dragão, Jardim do Morro, Aliados and Bolhão, and also the utility places as Hospital São João and IPO Metro Station and Universities.

The representation of the results from different days throughout the year will be on appendix. Each day will be a representative of each season of the year.

Location	Common/Sometimes
Maia	Common
Senhora Da Hora Metro Station	Common
Norte Shopping	Common
Hospital São João	Common
IPO Metro Station and Universities	Common
Areosa	Common
Cordoaria	Common
Aliados and Bolhão	Common
Trindade	Common
Rotunda da Boavista	Common
Marquês	Common
Carolina Michaelis Metro Station	Common
Campo 24 de Agosto Metro Station	Common
Jardim do Morro	Common
General Torres	Sometimes
El Corte Inglés	Common
Santo Ovidio	Common
Estádio do Dragão	Common
Campanhã	Common
Francisco Sá Carneiro Airport	Sometimes
Matosinhos	Sometimes

Table 5.6: Most Used Areas of the Network In Different Seasons

The list is almost equal to the January week list results. The only difference is that Matosinhos. Matosinhos only appear in the summer season since it is the area near to the sea and people tend to go to the beach in the summer. So, Matosinhos is the main difference in the results with his appearance on the list of most used areas because of this area appearance at the summer season.

5.2.2 Ticket Type Results

Here we are going to show which type of ticket is used the most. Our intention is to evaluate the use of each type of ticket existent and see if something could be done to better suit the passenger's needs.

Table 5.18 shows each type of ticket with their correspondent number of tickets and the ratio of tickets that are from that type. We made a table for each month. This way we can see the impact of each month on the results. The rest of the results are in appendix.

Ticket Type	Number Of Tickets	Ratio of tickets
Z2	1811170	66.924387
Z3	462505	17.089982
T1	135233	4.996983
Z6	49356	1.823749
T2	24636	0.910323
T3	4482	0.165614
Z4	151816	5.609740
Z32	20930	0.773383
Z5	38585	1.425751
Rede Geral(ABC)	6435	0.237779
Z1	433	0.016000
Z7	301	0.011122
C1-6 C8-C11 C16 N1 N10-11 N16 S1-2 S7-9	14	0.000517
Z12	2	0.000074
Z8	271	0.010014
Z9	124	0.004582

Figure 5.18: Ticket Type Results of January

The ticket type results show that the ticket type "Z2" is the most used type with is the lowest ratio value being 61 per cent and is highest 70 per cent. Following the "Z2" ticket type is the "Z3" where is the lowest ratio value was 17 and the highest 21. This means that "Z2" and "Z3" tickets together represent almost 90 Percent of the all tickets used throughout all months of the year. It also represents that passengers use occasional tickets to make short journeys because they only go through 2 or 3 zones.

5.3 Spatio-Temporal Results

In this section, we will show the results of the approaches presented in section 4.6. The first objective was to find the space and time profiles of the transport network utilization. The second as objective was to find the areas where trip-chain happens.

5.3.1 K-Means Results

The use of K-Means to profile the users of the network in terms of space and time was done by day. So, we will show the results for every day of a week. Then we will describe and compare those results. Afterwards, the same will be done to the results for different seasons of the year. For the week results, January 14 to 18 week was chosen(Figures 5.19 to 5.23). For the seasons results from the days of January 15(Figure 5.20), April 9(Figure A.19), July 9(Figure A.20) and November 5(Figure A.21) were chosen and represented on appendix.

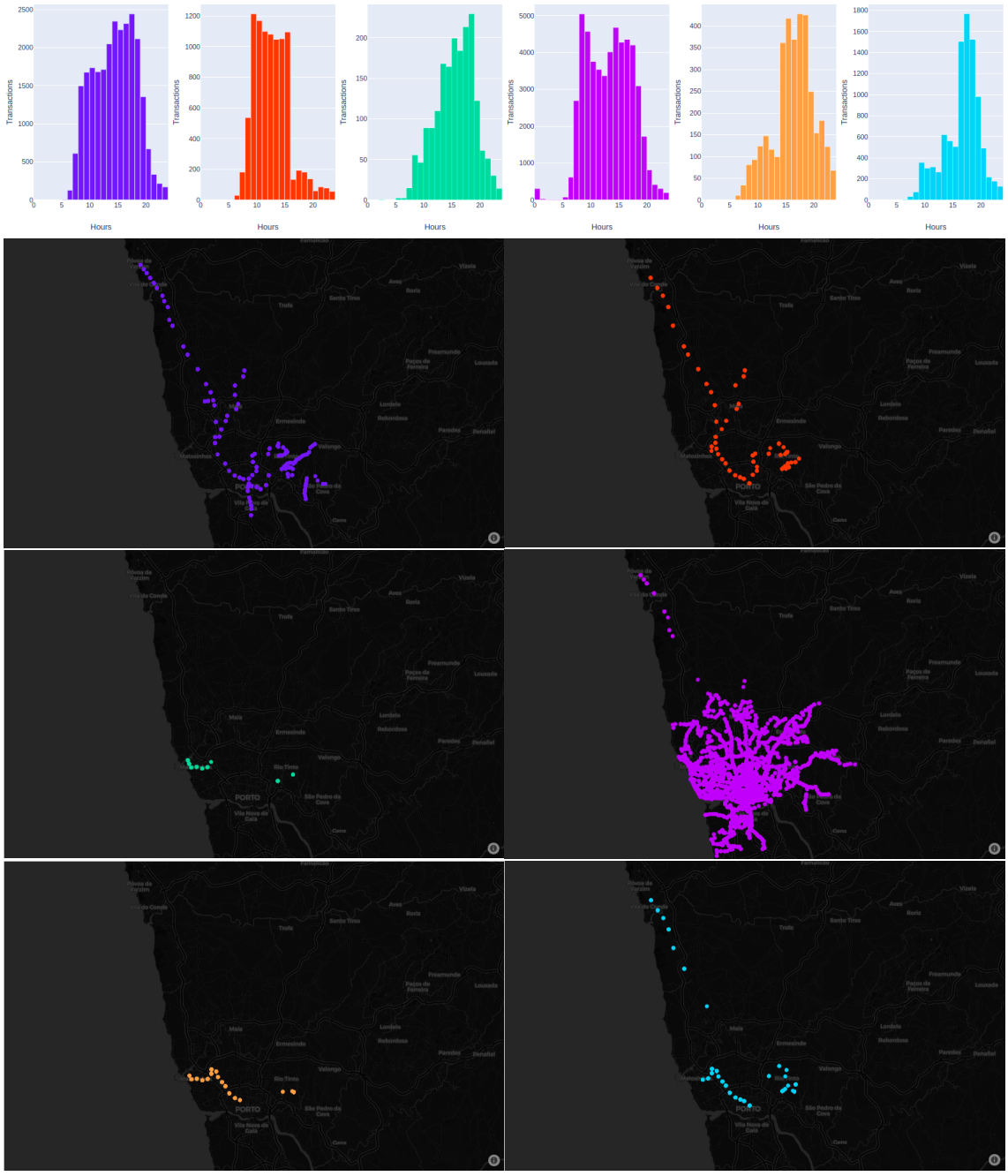


Figure 5.19: January 14 Spatio-Temporal Patterns

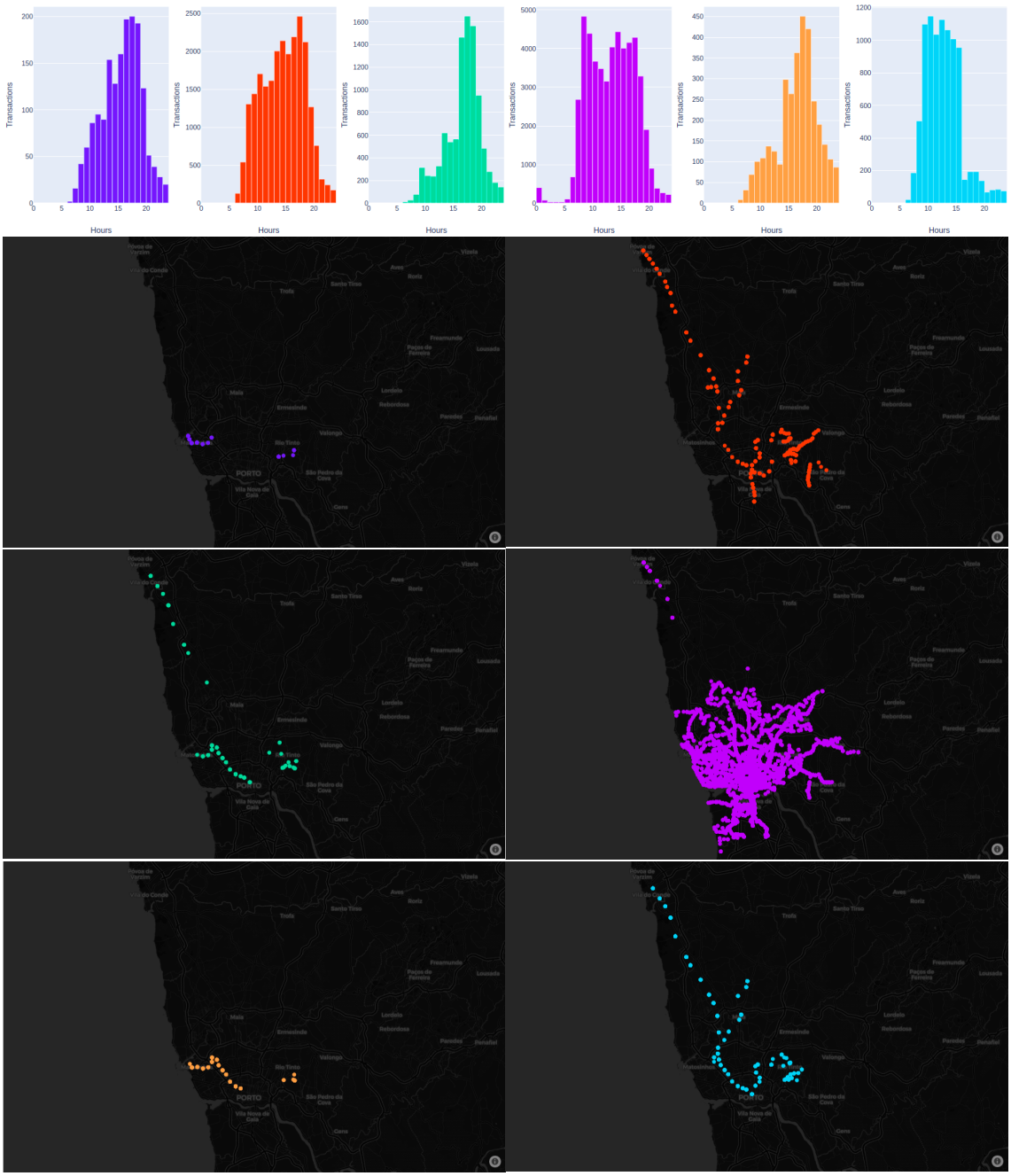


Figure 5.20: January 15 Spatio-Temporal Patterns

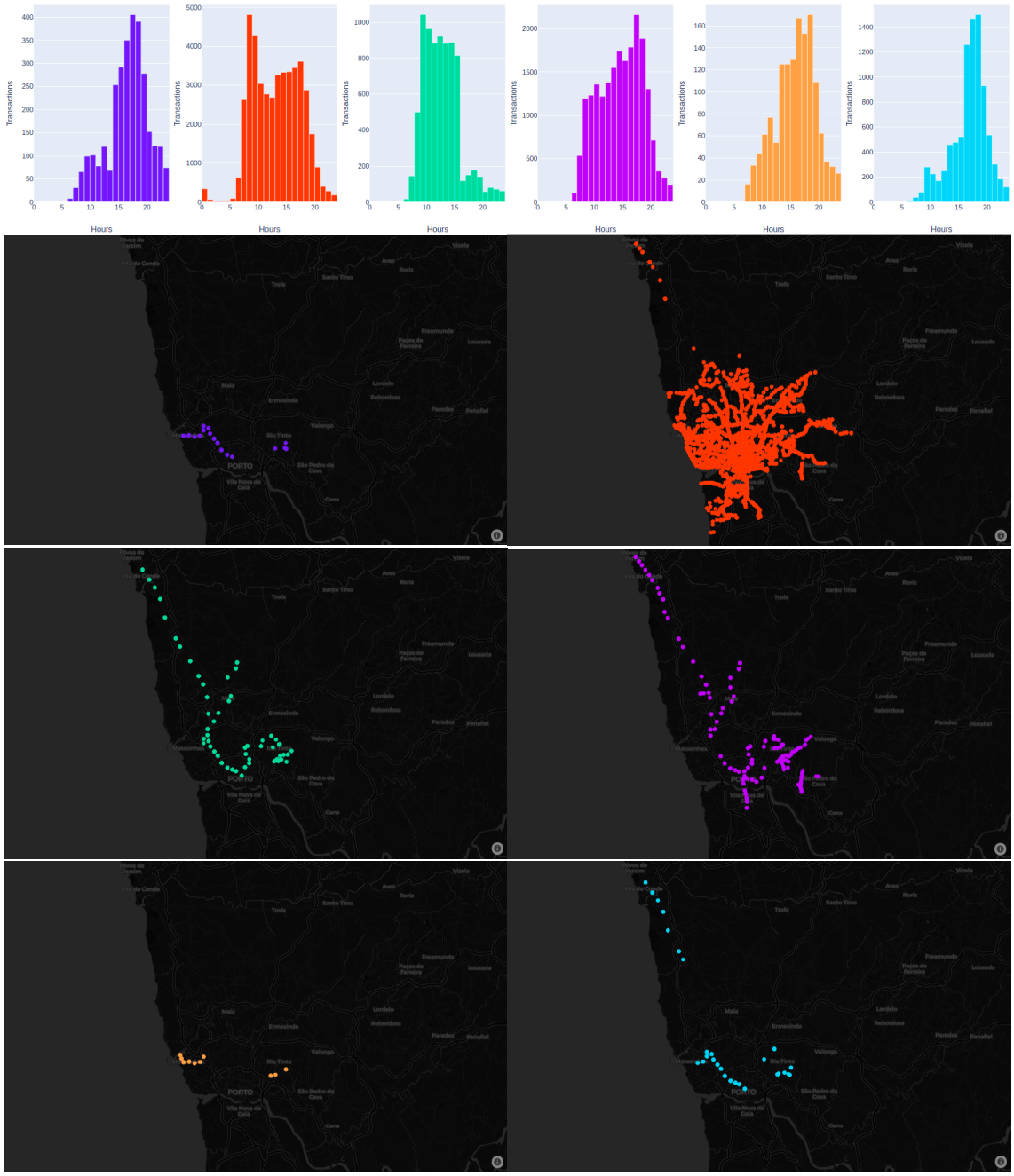


Figure 5.21: January 16 Spatio-Temporal Patterns

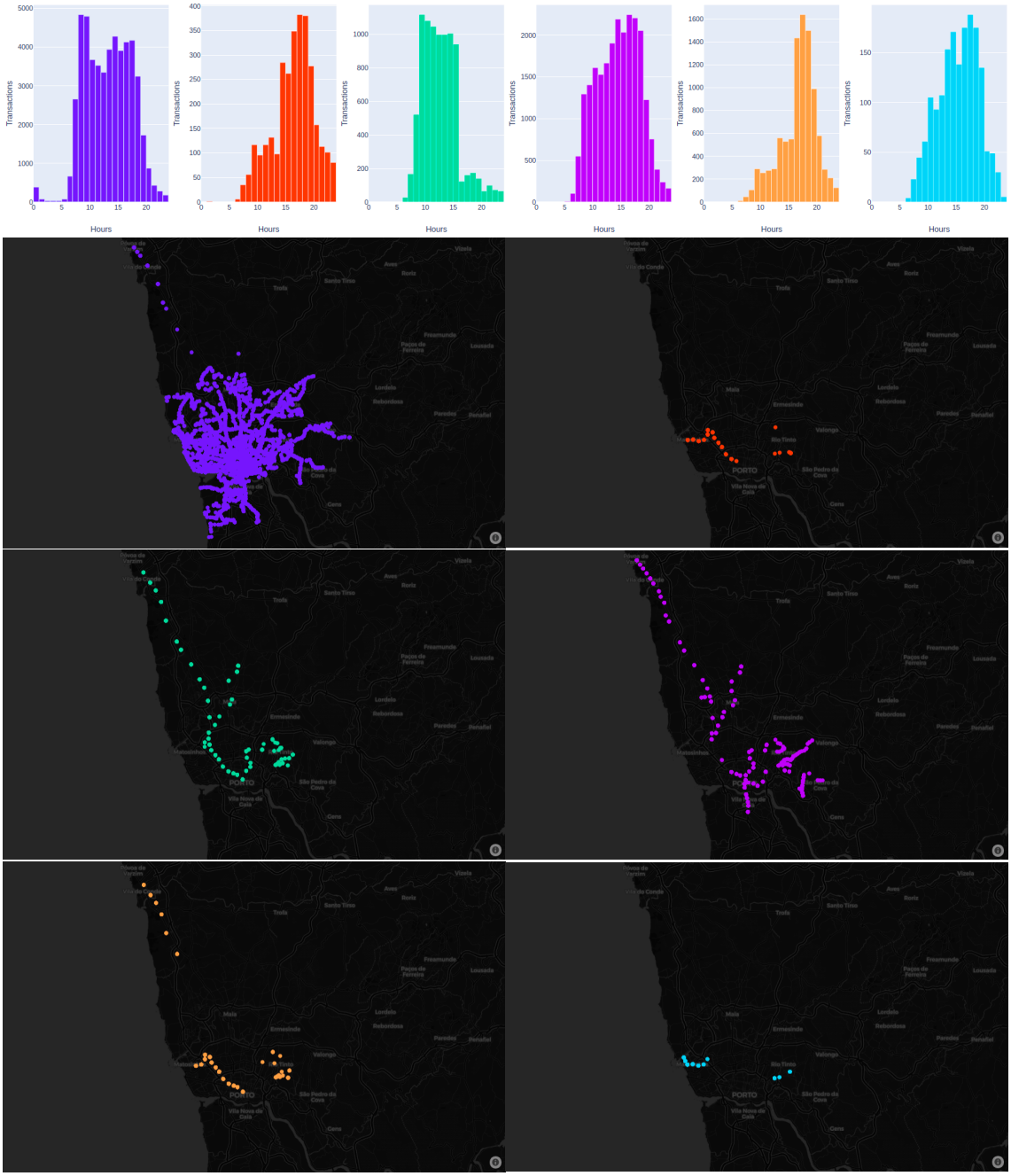


Figure 5.22: January 17 Spatio-Temporal Patterns

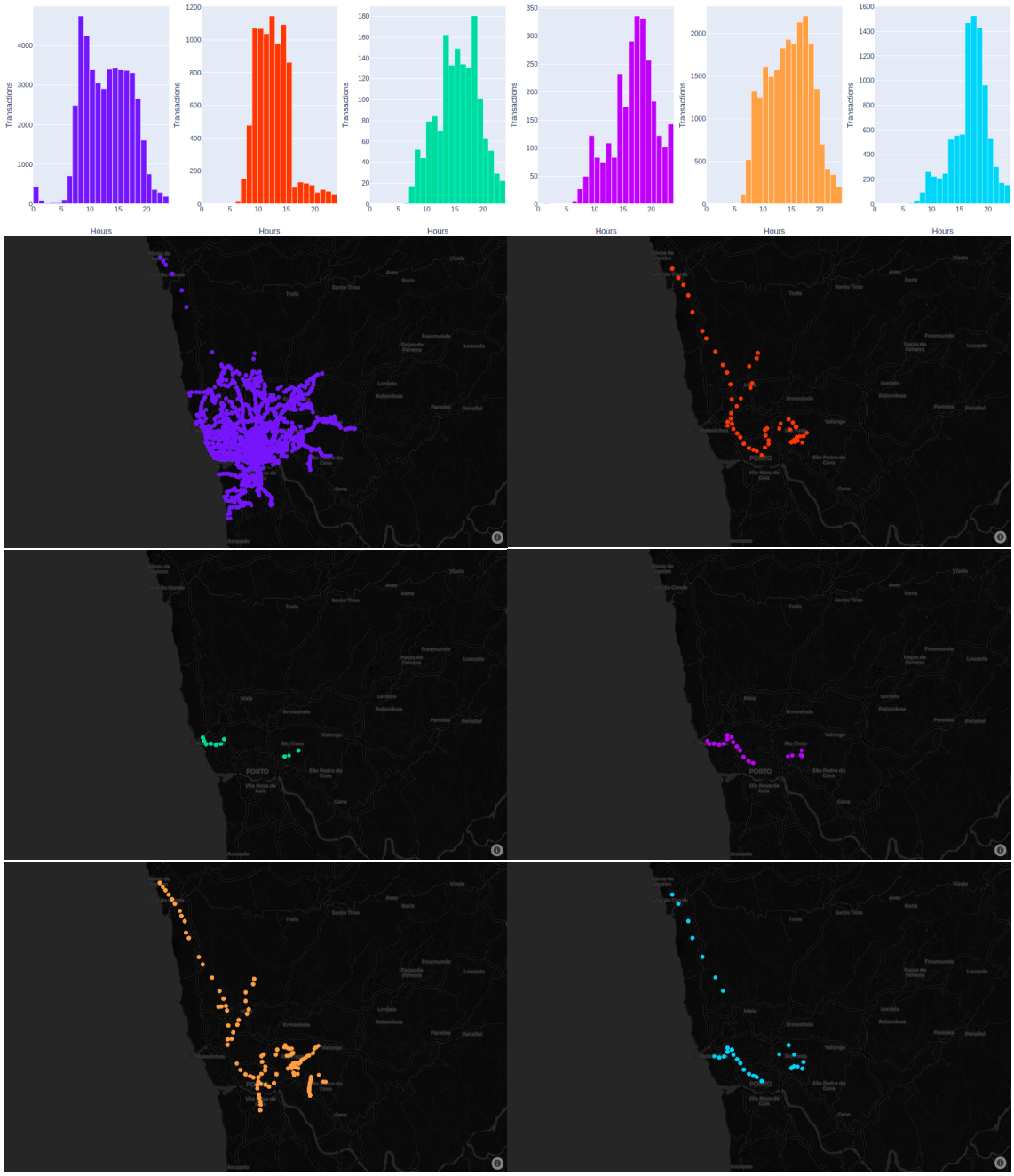


Figure 5.23: January 18 Spatio-Temporal Patterns

The results show that throughout the January week and the days that represent the different seasons of the year, there are always the same 6 clusters. From those clusters, 5 clusters show a profile of utilization and 1 that is probably agglomerating the rest of the entries that don't have a profile. This last cluster normally represents a big part of the network. These results show the same spatio-temporal patterns throughout the year. Also, there are some locations that belong to more than one cluster but those clusters have few differences between themselves. This also

happens because the locations are not unique. Each entry has a location and an hour so we are making profiles of the utilization of the network and not profiles of users utilization of the network. The clusters show 3 different patterns in terms of time. Those patterns are the morning usage, the afternoon usage and the morning and afternoon usage of the network. There are always 3 clusters with afternoon usage, 2 with the morning and afternoon usage and one with the morning. For example, in Figure A.21 the clusters green, yellow and blue are the 3 clusters with the afternoon usage. The purple and pink cluster are the ones with the morning and afternoon usage. The red cluster is the one with the morning usage.

5.3.2 Trip-Chain Results

In the trip-chain results, we only select the pairs that had at least 500 transactions. This way we can be more sure that these are significant results. We will show the trip-chain results of each day of a week. The week chosen was February 4 to 8.

0	1	counter
39	42	1283
42	39	771
42	42	2084
75	42	1086
81	42	549
88	42	672
89	42	776

(a) February 4 Trip-chain Pairs

0	1	counter
39	42	1265
42	39	695
42	42	1933
75	42	1001
81	42	529
88	42	697
89	42	737

(b) February 5 Trip-chain Pairs

0	1	counter
39	42	1147
42	39	693
42	42	1931
75	42	925
81	42	532
88	42	673
89	42	670

(c) February 6 Trip-chain Pairs

0	1	counter
39	42	1206
42	39	655
42	42	2017
75	42	1054
81	42	519
88	42	665
89	42	695

(d) February 7 Trip-chain Pairs

0	1	counter
39	42	1197
42	39	697
42	42	2172
75	42	1122
81	42	553
88	42	737
89	42	730

(e) February 8 Trip-chain Pairs

Figure 5.24: Trip-Chain Pairs with more than 500 Transactions

The next table and figure will expose the color and location of each zone id of the February week results.

Zone ID	Location	Color
39	Rotunda da Boavista	Yellow
42	Downtown Porto	Grey
75	Hospital/Universities	Dark Green
81	Marquês	Light Green
88	El Corte Inglés	White
89	Santo Ovidio	Purple

Table 5.7: Zones ID, location and respective color

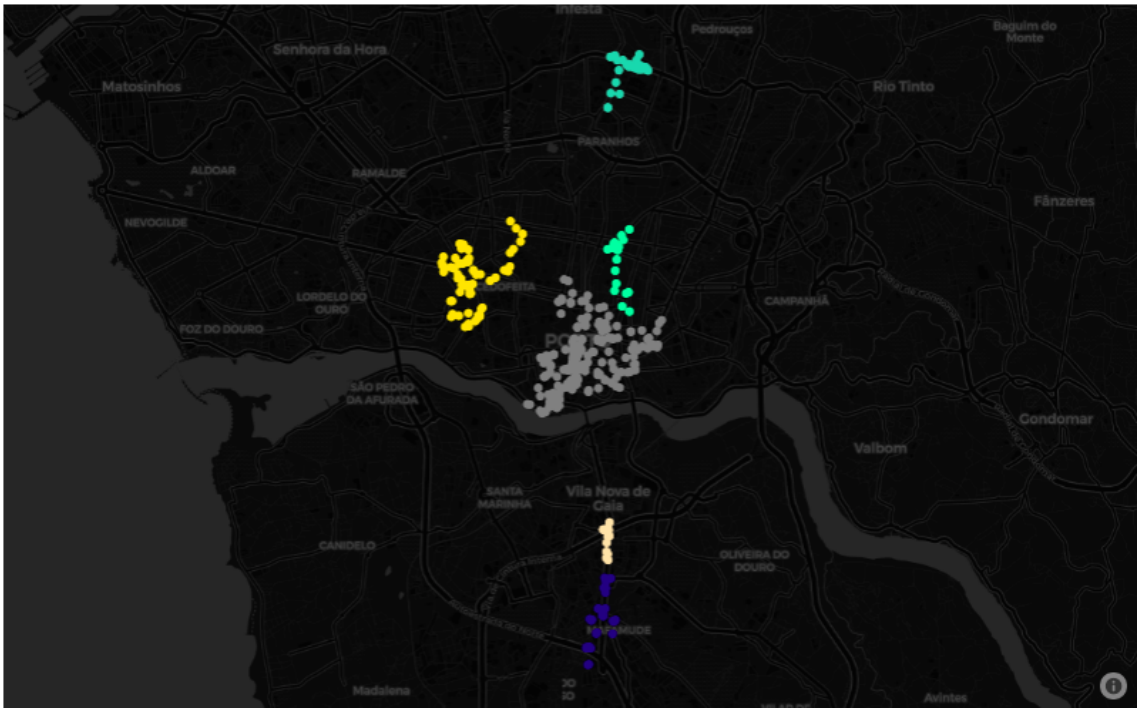


Figure 5.25: Trip-Chain Zones of Figure 5.24

The results show that the most common trip-chains are:

- Rotunda da Boavista to Dowtown Porto
- Downtown Porto to Rotunda da Boavista
- Downtown Porto to Downtown Porto

- Hospital/Universities to Downtown Porto
- Marquês to Downtown Porto
- El Corte Inglés to Downtown Porto
- Santo Ovidio to Downtown Porto

Chapter 6

Conclusion and Future Work

6.1 Future Work

As future work, I would look for some new approaches with the same objective as the ones done in this study. The first approach that I would do is similar to the SpatioTemporal where we used the K-means. In that approach, we found some of the SpatioTemporal Patterns of the network, but those patterns aren't a profile of users utilization, they are profiles of the network utilization. I would improve it by profiling the user's utilization of the network. That way we could find the existent users profiles of the network and gather more information about Spatio-Temporal Patterns. Another approach that I would explore is a way to find a different type of temporal profiles of users. In our approach we profiled the users by all hours of the day, and instead of doing it, I would improve it by dividing the hours of the day into groups of hours, for example, 6am to 9 am, 10 am to 1 pm, etc. This way we would mitigate the diffuse temporal patterns that happens sometimes in the non-frequent passengers data. A further approach that I would do is a Trip-Chain model. Our results on the Trip-Chain of Zones show that Trip-Chain exists in the non-frequent passengers, so I would try to do a Trip-Chain model of zones as we did. This way we would have more instances to train and less possible results making it easier to work.

6.2 Conclusion

Understanding non-frequent public transport passengers through data mining was our main objective and for that, we came up with 3 different types of patterns that these passengers could have. Those patterns were the Temporal Patterns, the Spatial Patterns and the SpatioTemporal Patterns. From those patterns, we then proceed to some approaches to retrieve knowledge in terms of those patterns. We found several passenger Temporal Profiles of usage of the network and that the transport network has a consistent flow of new passengers/tickets entering in. So, in Temporal Patterns, we found that the utilization of the transport network is constant throughout the day and that the average of new non-frequent passengers by day through the year is 15.7%. Also, the non-frequent passengers have some areas of the network that they use more than others. Those areas that are

used the most are the change of line or transport areas, the shopping's areas, the touristic areas, and the utility places as Hospital and Universities. They also use more certain types of tickets than others. There are two main types of tickets correspond to 90 per cent of all the tickets used in the transport network and those types of tickets show that the journeys made by the non-frequent passengers are short, from 2 to 3 zones. Further, we detected that Trip-Chain exists in non-frequent passengers, at least in terms of zones, and that the network as some profiles in terms of Space and Time. Where different areas of the network have different temporal usage. These findings show that non-frequent passengers data represent knowledge of the transport network in all possible terms of patterns, the Temporal, the Spatial and the SpatialTemporal Patterns. Consequently, we recommend looking also to the non-frequent passenger's data as they also hold patterns in their data.

Appendix A

Results

A.1 New Passengers Results

1	9015	0.196675
2	8751	0.158092
3	7991	0.154121
4	7928	0.147853
5	8418	0.149470
6	6808	0.195189
7	7419	0.280545
8	8778	0.161088
9	7052	0.139699
10	6937	0.135942
11	6657	0.137448
12	8332	0.149129
13	6844	0.192545
14	4855	0.188756
15	7757	0.147842
16	7429	0.140212
17	7458	0.139747
18	7611	0.140877
19	8852	0.151861
20	6989	0.191485
21	5055	0.200540
22	7831	0.148955

1	4764	0.240485
2	6672	0.189207
3	6536	0.181611
4	7849	0.258710
5	10853	0.322430
6	6853	0.198643
7	12257	0.246204
8	6936	0.185932
9	6391	0.182778
10	8148	0.200937
11	9907	0.270927
12	5064	0.235382
13	5376	0.179733
14	4820	0.164303
15	4861	0.165183
16	4518	0.160088
17	5423	0.173226
18	5251	0.233710
19	5818	0.271248
20	5467	0.173478
21	5771	0.182841
22	5038	0.165077
23	5327	0.169547

1	10608	0.233066
2	6506	0.212607
3	10531	0.177355
4	8434	0.148729
5	7883	0.141156
6	7540	0.137629
7	8202	0.140667
8	6344	0.184805
9	5102	0.220971
10	11432	0.342368
11	7807	0.129396
12	7263	0.134760
13	7402	0.135672
14	8297	0.145329
15	7506	0.203884
16	4516	0.200000
17	7183	0.144836
18	7032	0.135632
19	7145	0.135769
20	7230	0.132704
21	8065	0.140442
22	7273	0.192326

1	11414	0.174153
2	9195	0.151770
3	8821	0.144122
4	8438	0.140755
5	8589	0.143499
6	8081	0.206712
7	5252	0.185872
8	8260	0.148925
9	8157	0.144259
10	8439	0.140185
11	8465	0.141060
12	8374	0.140463
13	6271	0.174709
14	4822	0.183102
15	8804	0.154915
16	8241	0.142874
17	8423	0.142353
18	8515	0.145270
19	8440	0.144915
20	6708	0.181858
21	4839	0.190280
22	8255	0.153684
23	8157	0.144420
24	8122	0.141343
25	8055	0.142771
26	7928	0.138941
27	5994	0.176502
28	6393	0.227485
29	7729	0.145663
30	7836	0.141044
31	8091	0.144464

(a) July New Passengers

1	7291	0.196544
2	6383	0.181604
3	5014	0.227413
4	4449	0.258093
5	6547	0.207098
6	6931	0.218169
7	6914	0.221262
8	6595	0.206468
9	6773	0.203620
10	5368	0.253148
11	4615	0.274653
12	6875	0.228846
13	6823	0.224065
14	6679	0.218026
15	5490	0.287615
16	6880	0.242681
17	5092	0.258490
18	4401	0.275838
19	6595	0.227367
20	6707	0.225802
21	6340	0.215727
22	6137	0.213201
23	6026	0.202255
24	4562	0.232104
25	5509	0.279404
26	5796	0.196848
27	5734	0.188148
28	5417	0.177357
29	5375	0.174746
30	5415	0.171031
31	4775	0.223475

(b) August New Passengers

1	6974	0.237566
2	10182	0.177932
3	8669	0.157381
4	8589	0.154723
5	8028	0.150438
6	8343	0.148820
7	6282	0.179681
8	4858	0.193392
9	8540	0.151362
10	8481	0.143919
11	8296	0.141858
12	8061	0.136424
13	8652	0.141876
14	8156	0.184183
15	5255	0.183651
16	8814	0.140830
17	7667	0.125235
18	7101	0.117953
19	7463	0.120567
20	7685	0.120869
21	6165	0.168678
22	4385	0.170060
23	7556	0.120682
24	6757	0.107781
25	6444	0.105243
26	6962	0.108974
27	7659	0.124953
28	5436	0.163028
29	3712	0.164226
30	6882	0.110975

(c) September New Passengers

1	10901	0.147412
2	7340	0.113889
3	6786	0.112123
4	7292	0.116603
5	5754	0.156840
6	4230	0.159076
7	6482	0.114148
8	6211	0.110122
9	6173	0.110671
10	6222	0.108141
11	6969	0.118063
12	6371	0.169916
13	3839	0.163710
14	5763	0.115228
15	5341	0.106764
16	5461	0.105181
17	5768	0.107582
18	6310	0.114617
19	6007	0.169977
20	3769	0.160185
21	6078	0.118010
22	6621	0.120240
23	5464	0.102247
24	5136	0.106256
25	6543	0.113738
26	5410	0.151745
27	6155	0.184447
28	5521	0.106599
29	5251	0.098620
30	5125	0.094281
31	5851	0.103127

(a) October New Passengers

1	8329	0.131800
2	4734	0.153497
3	3772	0.151748
4	5149	0.103290
5	4644	0.093860
6	4497	0.091466
7	4025	0.108144
8	5387	0.107259
9	4683	0.146252
10	3054	0.137599
11	5177	0.099877
12	4550	0.090170
13	4495	0.089022
14	4507	0.089588
15	5155	0.097316
16	4604	0.137408
17	2864	0.130544
18	4615	0.095053
19	3986	0.083655
20	4004	0.082034
21	4339	0.089374
22	5212	0.097268
23	4513	0.130683
24	2756	0.124139
25	4419	0.089493
26	3957	0.120226
27	4206	0.083675
28	4082	0.082230
29	4918	0.090085
30	4459	0.128077

(b) November New Passengers

1	3470	0.135004
2	5442	0.096867
3	4314	0.083453
4	4287	0.082517
5	4236	0.082781
6	5597	0.101533
7	5939	0.150545
8	3085	0.129698
9	4189	0.084128
10	3933	0.076650
11	3932	0.075841
12	3907	0.076114
13	4677	0.088543
14	4580	0.120523
15	4494	0.154221
16	4597	0.086759
17	4355	0.084627
18	4683	0.091304
19	4620	0.090938
20	6701	0.105229
21	5191	0.126814
22	3429	0.141024
23	6529	0.121958
24	1922	0.118995
25	970	0.167820
26	4702	0.133360
27	5512	0.133424
28	4705	0.153572
29	3608	0.171949
30	6094	0.135482
31	6115	0.165449

(c) December New Passengers

A.2 Most Used Areas Results

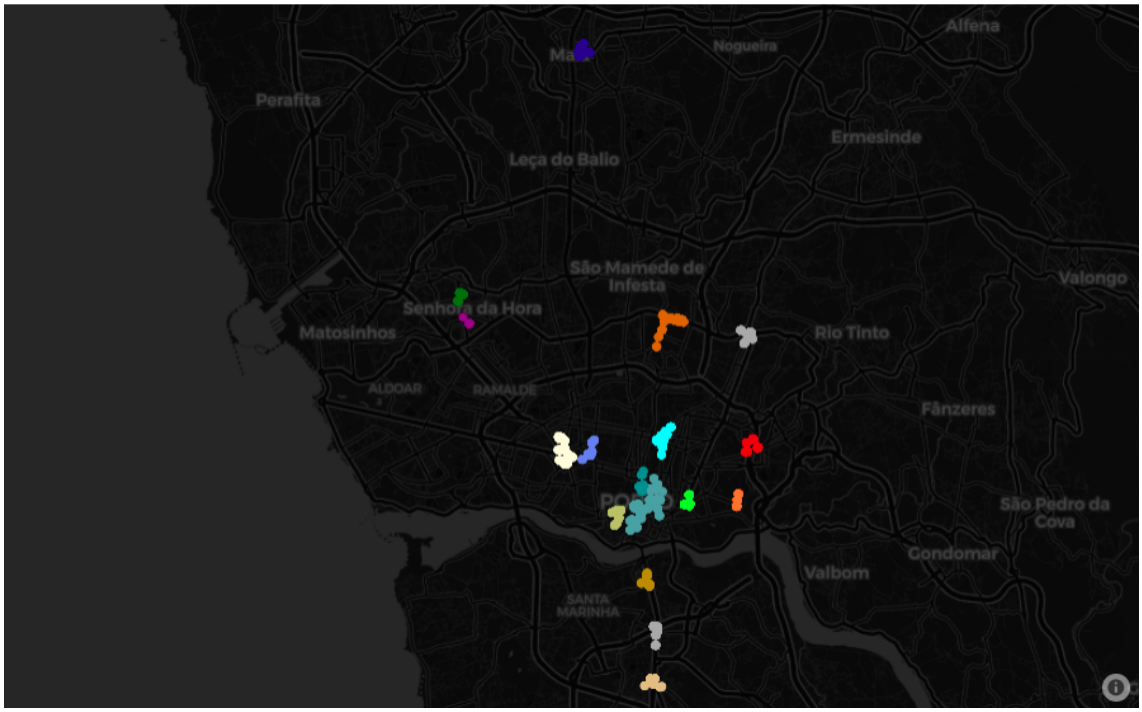


Figure A.4: January 8 DBSCAN Results

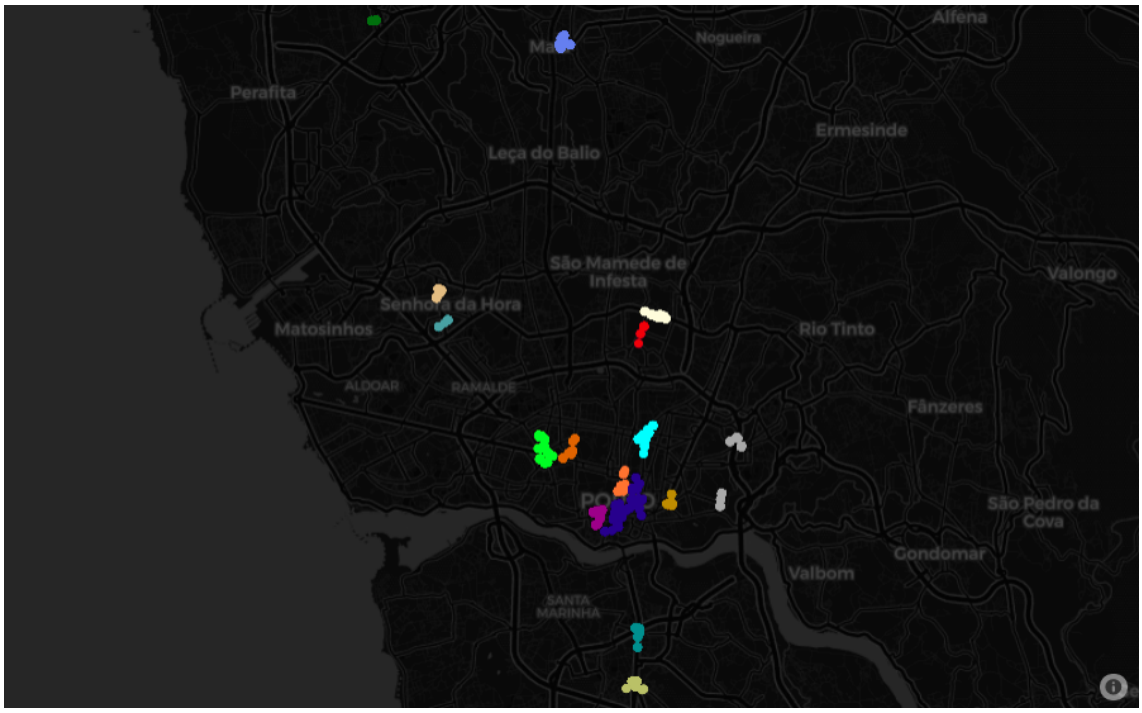


Figure A.5: April 9 DBSCAN Results - 17 Clusters

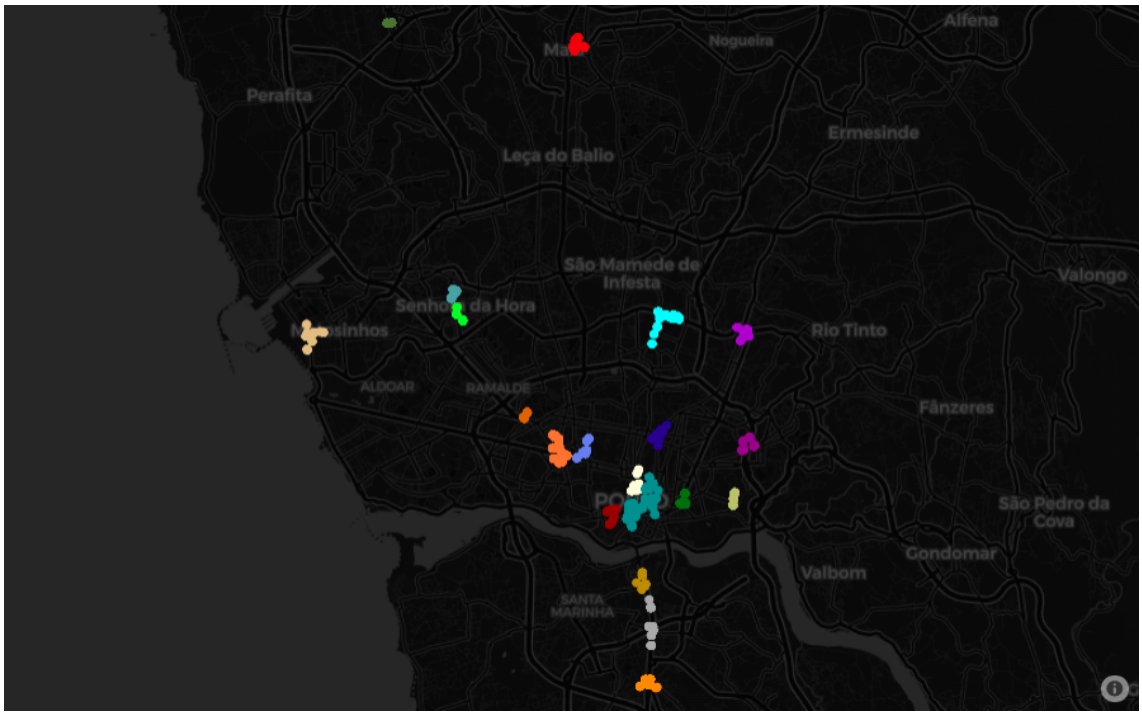


Figure A.6: July 9 DBSCAN Results - 21 Clusters

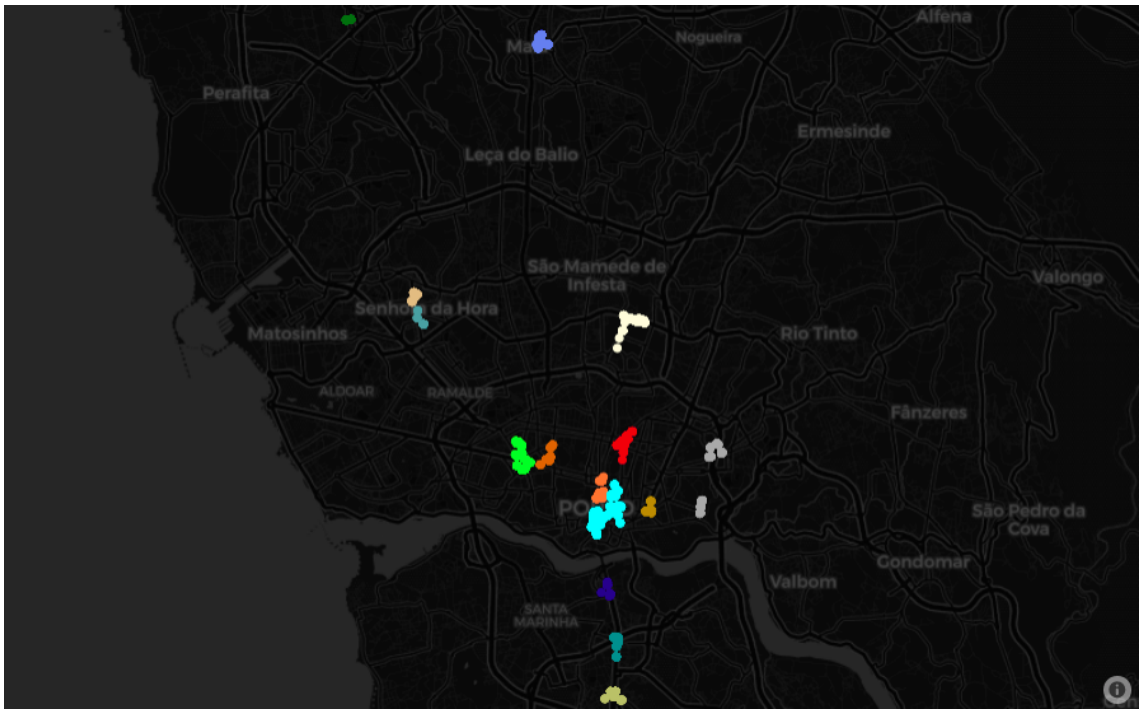


Figure A.7: November 5 DBSCAN Results - 16 Clusters

A.3 Ticket Type Results

Ticket Type	Number Of Tickets	Ratio of tickets
Z2	1737387	69.915195
Z3	445394	17.923357
Z4	151263	6.087062
Rede Geral(ABC)	7916	0.318552
Z5	36349	1.462741
Z32	22605	0.909661
T1	28553	1.149018
T3	959	0.038592
T2	5767	0.232073
Z7	258	0.010382
Z8	302	0.012153
Z6	48085	1.935016
Z9	89	0.003582
C1-6 C8-C11 C16 N1 N10-11 N16 S1-2 S7-9	23	0.000926
Z1	40	0.001610
C1-2 C6	1	0.000040
C1 S8	1	0.000040

Figure A.8: Ticket Type Results of February

Ticket Type	Number Of Tickets	Ratio of tickets
Z2	1802733	69.991462
Z3	474084	18.406404
T1	17355	0.673811
Z4	159131	6.178292
Z32	33034	1.282552
Z5	36556	1.419294
T3	682	0.026479
T2	3578	0.138917
Z6	47762	1.854369
Z7	192	0.007454
C1-6 C8-C11 C16 N1 N10-11 N16 S1-2 S7-9	8	0.000311
Z8	355	0.013783
Z9	131	0.005086
Z1	44	0.001708
Z16	2	0.000078

Figure A.9: Ticket Type Results of March

Ticket Type	Number Of Tickets	Ratio of tickets
Z3	519392	18.674970
Z2	1913471	68.799699
Z32	49761	1.789179
Z4	189627	6.818123
T1	13633	0.490181
Z6	51300	1.844514
Z5	39864	1.433328
Z9	145	0.005214
T2	2819	0.101358
Z8	286	0.010283
Z7	302	0.010859
C1-6 C8-C11 C16 N1 N10-11 N16 S1-2 S7-9	21	0.000755
T3	565	0.020315
Z1	34	0.001222

Figure A.10: Ticket Type Results of April

Ticket Type	Number Of Tickets	Ratio of tickets
Z2	1126995	64.322123
Z3	384203	21.928006
Z6	33552	1.914947
Z32	34009	1.941030
Z4	136603	7.796481
Z5	35731	2.039311
T1	345	0.019691
T2	96	0.005479
Z9	96	0.005479
Z8	282	0.016095
Z7	174	0.009931
C1-6 C8-C11 C16 N1 N10-11 N16 S1-2 S7-9	10	0.000571
T3	14	0.000799
Z1	1	0.000057

Figure A.11: Ticket Type Results of May

Ticket Type	Number Of Tickets	Ratio of tickets
Z2	1893180	68.004280
Z3	543759	19.532184
Z32	50021	1.796787
Z4	190365	6.838037
Z5	40404	1.451338
Z6	55761	2.002972
T3	323	0.011602
T1	7503	0.269513
T2	1624	0.058335
Z8	415	0.014907
Z9	106	0.003808
Z7	410	0.014727
C1-6 C8-C11 C16 N1 N10-11 N16 S1-2 S7-9	5	0.000180
Z1	37	0.001329

Figure A.12: Ticket Type Results of June

Ticket Type	Number Of Tickets	Ratio of tickets
Z2	2181304	67.344006
T2	1692	0.052238
Z3	644728	19.904867
Z32	55004	1.698154
Z6	70376	2.172738
Z4	225064	6.948464
Z5	51752	1.597754
T1	7736	0.238836
Z8	366	0.011300
Z7	421	0.012998
Z9	143	0.004415
C1-6 C8-C11 C16 N1 N10-11 N16 S1-2 S7-9	6	0.000185
Z12	3	0.000093
Z16	9	0.000278
T3	377	0.011639
Z1	65	0.002007
Rede Geral(ABC)	1	0.000031

Figure A.13: Ticket Type Results of July

Ticket Type	Number Of Tickets	Ratio of tickets
Z4	136860	8.816657
Z32	33937	2.186255
Z2	961475	61.939175
Z3	338234	21.789370
Z1	2	0.000129
Z5	38604	2.486908
Z6	42190	2.717922
T1	121	0.007795
T2	43	0.002770
T3	9	0.000580
Z9	183	0.011789
Z8	373	0.024029
Z7	258	0.016621

Figure A.14: Ticket Type Results of August

Ticket Type	Number Of Tickets	Ratio of tickets
Z2	2123899	68.216284
Z4	222782	7.155406
Z3	586722	18.844585
Z5	49905	1.602870
Z32	55502	1.782637
Z6	66796	2.145382
T1	5106	0.163997
T2	1122	0.036037
T3	331	0.010631
Z8	584	0.018757
Z9	137	0.004400
Z7	510	0.016380
C1-6 C8-C11 C16 N1 N10-11 N16 S1-2 S7-9	50	0.001606
Z1	31	0.000996
S1-2 S8-9	1	0.000032

Figure A.15: Ticket Type Results of September

Ticket Type	Number Of Tickets	Ratio of tickets
Z2	2148547	69.954870
Z3	556186	18.108945
Z32	50234	1.635576
Z4	208605	6.792002
T1	4660	0.151726
Z5	43466	1.415216
Z6	56993	1.855644
T2	1095	0.035652
Z1	100	0.003256
Z9	176	0.005730
Z7	446	0.014521
T3	355	0.011558
Z8	440	0.014326
Z12	2	0.000065
C1-6 C8-C11 C16 N1 N10-11 N16 S1-2 S7-9	28	0.000912

Figure A.16: Ticket Type Results of October

Ticket Type	Number Of Tickets	Ratio of tickets
Z2	1785684	70.165614
Z3	475077	18.667395
Z6	48852	1.919562
Z32	28879	1.134754
Z4	162184	6.372762
Z5	38995	1.532247
T1	3395	0.133401
T3	259	0.010177
T2	812	0.031906
Z8	308	0.012102
Z7	335	0.013163
Z9	118	0.004637
C1-6 C8-C11 C16 N1 N10-11 N16 S1-2 S7-9	3	0.000118
Z1	53	0.002083
C1-2	1	0.000039
C1 C6 C8	1	0.000039

Figure A.17: Ticket Type Results of November

Ticket Type	Number Of Tickets	Ratio of tickets
Z2	1808896	70.672644
Z4	150777	5.890780
Z3	485451	18.966323
Z6	48688	1.902215
Z32	20261	0.791587
T3	331	0.012932
Z5	39338	1.536916
T1	3873	0.151316
Z9	102	0.003985
Z8	288	0.011252
Z7	276	0.010783
T2	1231	0.048095
Z1	29	0.001133
C1-6 C8-C11 C16 N1 N10-11 N16 S1-2 S7-9	1	0.000039

Figure A.18: Ticket Type Results of December

A.4 K-means Results

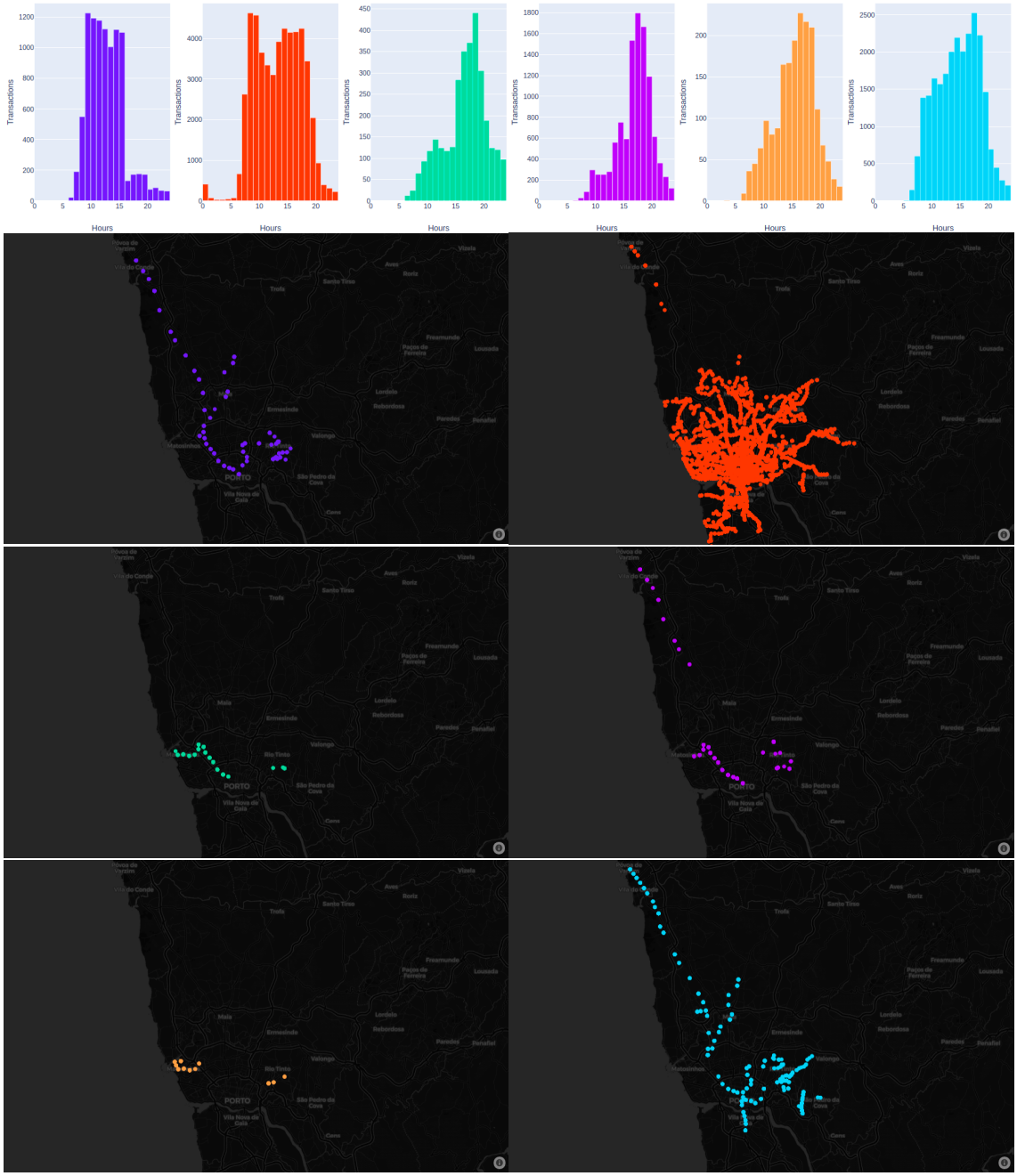


Figure A.19: April 9 Spatio-Temporal Patterns

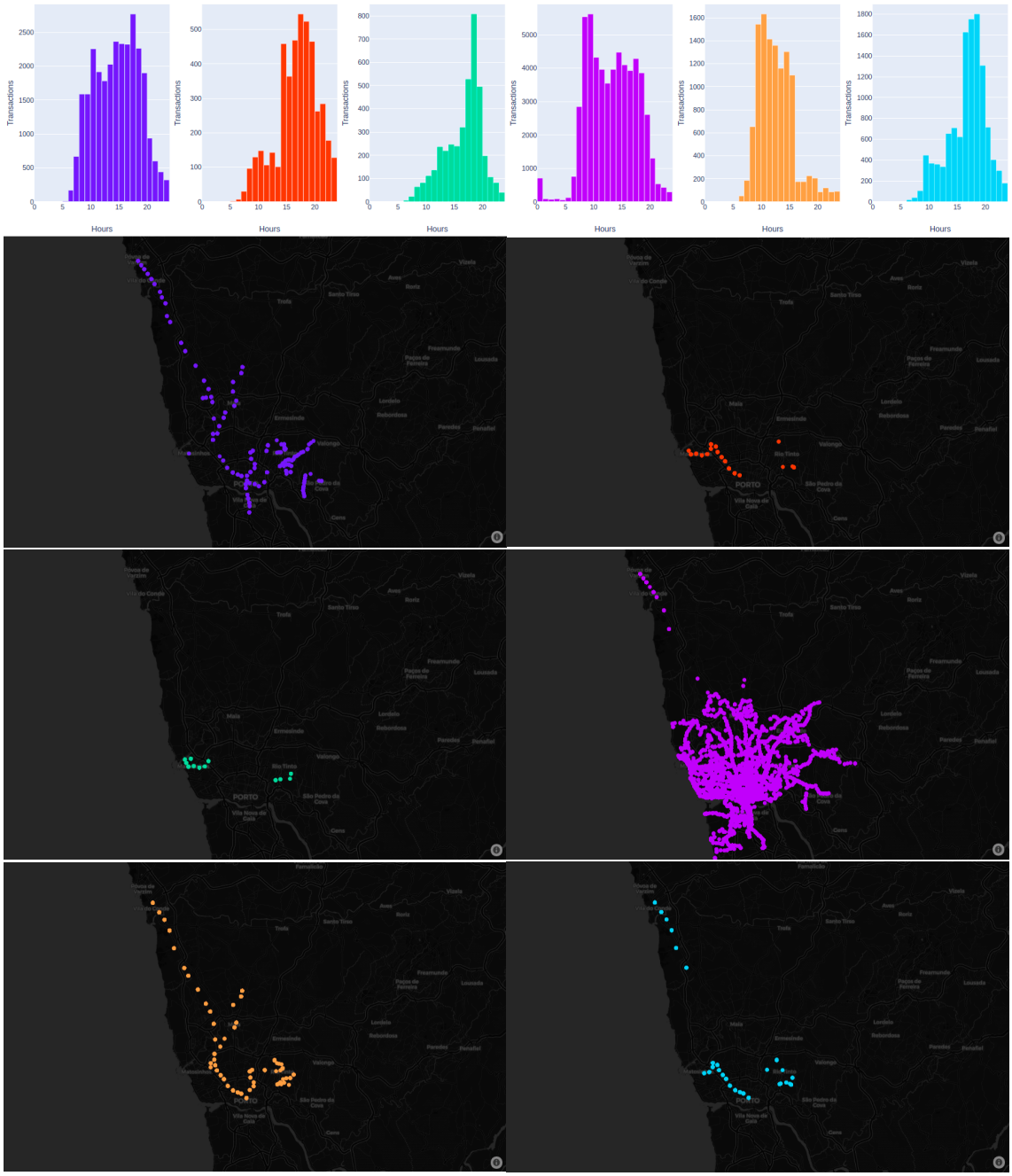


Figure A.20: July 9 Spatio-Temporal Patterns

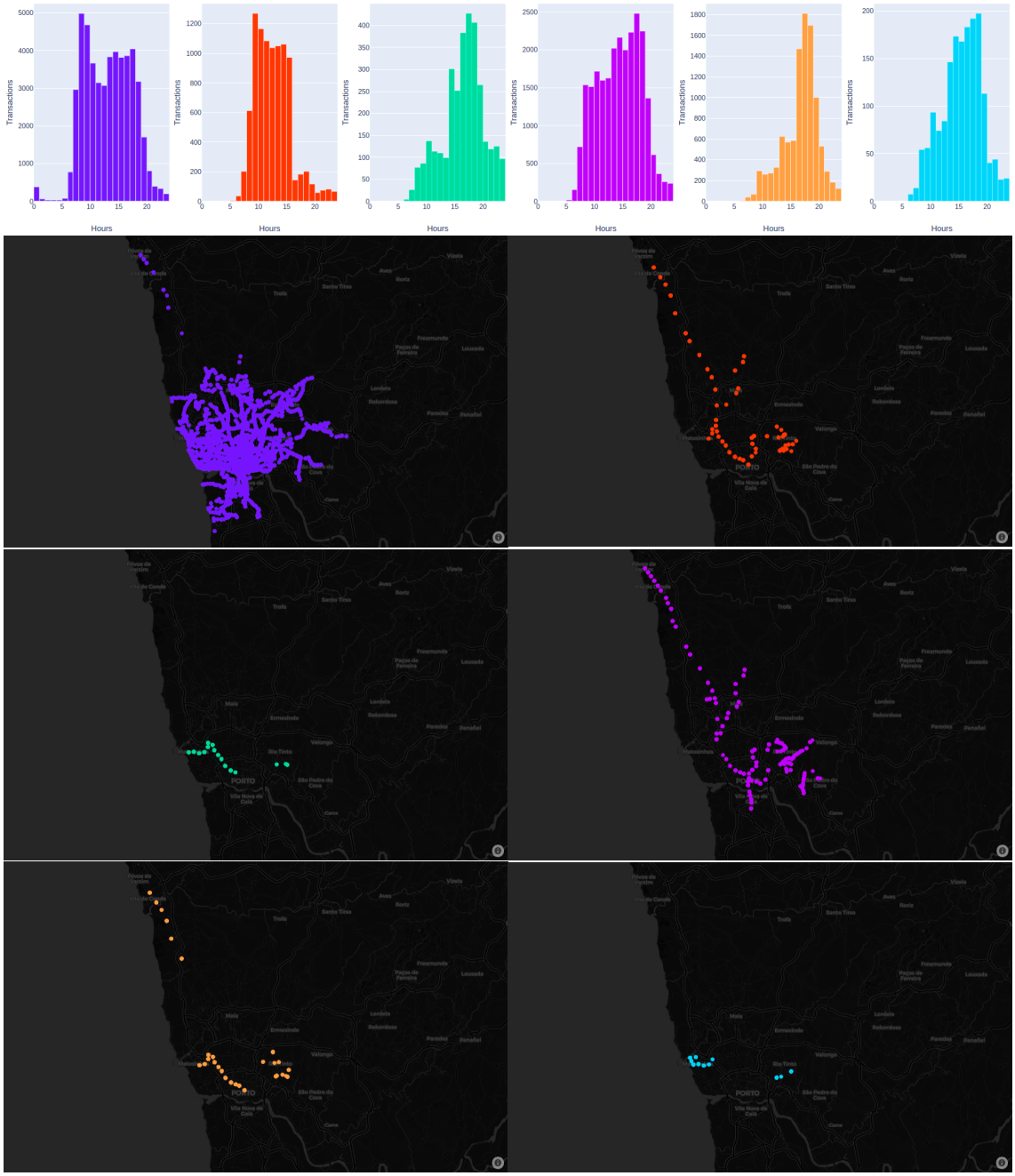


Figure A.21: November 5 Spatio-Temporal Patterns

Bibliography

[1] James J. Barry et al. “Origin and destination estimation in New York City with automated fare system data”. In: *Transportation Research Record* 1817 (2002), pp. 183–187. ISSN: 03611981. DOI: 10.3141/1817-24.

[2] Vincent D. Blondel et al. “Fast unfolding of communities in large networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* (2008). ISSN: 17425468. DOI: 10.1088/1742-5468/2008/10/P10008.

[3] James J. Barry, Robert Freimer, and Howard Slavin. “Use of entry-only automatic fare collection data to estimate linked transit trips in New York city”. In: *Transportation Research Record* 2112 (2009), pp. 53–61. ISSN: 03611981. DOI: 10.3141/2112-07.

[4] Qing Zhu, Yingzhe Wang, and Jiankou Li. “Public transport IC card data analysis and operation strategy research based on data mining technology”. In: *IFCSTA 2009 Proceedings - 2009 International Forum on Computer Science-Technology and Applications* 3 (2009), pp. 459–462. DOI: 10.1109/IFCSTA.2009.352.

[5] Jean Patrick Baudry, Cathy Maugis, and Bertrand Michel. “Slope heuristics: Overview and implementation”. In: *Statistics and Computing* 22.2 (2012), pp. 455–470. ISSN: 09603174. DOI: 10.1007/s11222-011-9236-1.

[6] Philip Sedgwick. *Pearson’s correlation coefficient*. 2012. DOI: 10.1136/bmj.e4483.

[7] Côme Etienne and Oukhellou Latifa. “Model-based count series clustering for bike sharing system usage mining: A case study with the vélib’ system of Paris”. In: *ACM Transactions on Intelligent Systems and Technology* 5.3 (2014), pp. 1–21. ISSN: 21576912. DOI: 10.1145/2560188.

[8] Anne Sarah Briand et al. “A mixture model clustering approach for temporal passenger pattern characterization in public transport”. In: *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015* (2015), pp. 1–10. ISSN: 2364-415X. DOI: 10.1109/DSAA.2015.7344847.

[9] Azalden Alsger et al. “Validating and improving public transport origin-destination estimation algorithm using smart card fare data”. In: *Transportation Research Part C: Emerging Technologies* (2016). ISSN: 0968090X. DOI: 10.1016/j.trc.2016.05.004.

[10] Gabriel Goulet Langlois, Haris N. Koutsopoulos, and Jinhua Zhao. “Inferring patterns in the multi-week activity sequences of public transport users”. In: *Transportation Research Part C: Emerging Technologies* (2016). ISSN: 0968090X. DOI: 10.1016/j.trc.2015.12.012.

[11] António A. Nunes, Teresa Galvão Dias, and João Falcão E Cunha. “Passenger journey destination estimation from automated fare collection system data using spatial validation”. In: *IEEE Transactions on Intelligent Transportation Systems* 17.1 (2016), pp. 133–142. ISSN: 15249050. DOI: 10.1109/TITS.2015.2464335.

[12] Mohamed K. El Mahrsi et al. “Clustering Smart Card Data for Urban Mobility Analysis”. In: *IEEE Transactions on Intelligent Transportation Systems* 18.3 (2017), pp. 712–728. ISSN: 15249050. DOI: 10.1109/TITS.2016.2600515.

[13] Juanjuan Zhao et al. “Spatio-Temporal Analysis of Passenger Travel Patterns in Massive Smart Card Data”. In: *IEEE Transactions on Intelligent Transportation Systems* 18.11 (2017), pp. 3135–3146. ISSN: 15249050. DOI: 10.1109/TITS.2017.2679179.

[14] Chengmei Liu, Chao Gao, and Yingchu Xin. “Measuring the Diversity and Dynamics of Mobility Patterns Using Smart Card Data”. In: *11th International Conference on Knowledge Science, Engineering and Management (KSEM)*. Vol. 1. Springer International Publishing, 2018, pp. 438–451. ISBN: 9783319992471. DOI: 10.1007/978-3-319-99247-1{_}39. URL: http://link.springer.com/10.1007/978-3-319-99247-1_39.

[15] Cristina Pronello, Davide Longhi, and Jean Baptiste Gaborieau. “Smart card data mining to analyze mobility patterns in suburban areas”. In: *Sustainability (Switzerland)* 10.10 (2018), pp. 1–21. ISSN: 20711050. DOI: 10.3390/su10103489.

[16] Andante. *"Quem Somos"*. URL: https://www.linhandante.com/quemsomos.asp.

[17] Guru99. *"Data Mining Tutorial: Process, Techniques, Tools, EXAMPLES"*. URL: https://www.guru99.com/data-mining-tutorial.html.

[18] Scikit-Learn. *"Gaussian Mixture Models"*. URL: https://scikit-learn.org/stable/modules/mixture.html#gaussian-mixture.

[19] Scikit-learn. *" Linear and Quadratic Discriminant Analysis"*. URL: https://scikit-learn.org/stable/modules/lda_qda.html#mathematical-formulation-of-the-lda-and-qda-classifiers.

[20] Scikit-learn. *"Comparing different clustering algorithms on toy datasets"*. URL: https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html#sphx-glr-auto-examples-cluster-plot-cluster-comparison-py.