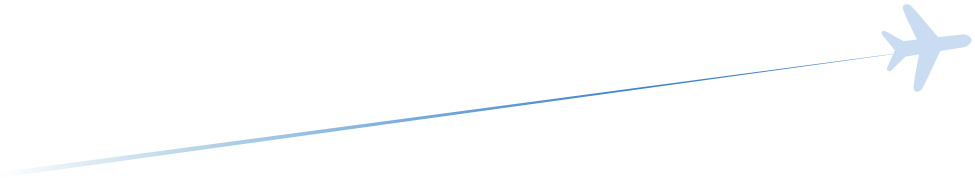
EXPLORATORY RESEARCH



Methodologies and Mobility Analytics Algorithms for the Analysis of the Door-to-Door Passenger Journey

Deliverable ID:	**D4.1**
Dissemination Level:	**PU**
Project Acronym:	**TRANSIT**
Grant:	**893209**
Call:	**H2020-SESAR-2019-2**
Topic:	**SESAR-ER4-10-2019**
Consortium Coordinator:	**Nommon**
Edition Date:	**21 May 2021**
Edition:	**00.01.00**
Template Edition:	**02.00.02**

Founding Members

EUROPEAN UNION EUROCONTROL

SESAR JOINT UNDERTAKING

Authoring & Approval

Authors of the document

Name/Beneficiary	Position/Title	Date
Alex Gregg (Nommon)	Deputy Project Coordinator	21/05/2021
Javier Burrieza (Nommon)	Project Team	21/05/2021
Rafael Jordá (Nommon)	Project Team	21/05/2021
Geoffrey Scozzaro (ENAC)	Project Team	21/05/2021
Clara Buire (ENAC)	Project Team	21/05/2021

Reviewers internal to the project

Name/Beneficiary	Position/Title	Date
Oliva García Cantú (Nommon)	Project Team	21/05/2021
Ricardo Herranz (Nommon)	Project Team	21/05/2021
Geoffrey Scozzaro (ENAC)	Project Team	21/05/2021
Clara Buire (ENAC)	Project Team	21/05/2021
Clarissa Livingston (ETH)	Project Team	21/05/2021
Stefano Penazzi (ETH)	Project Team	21/05/2021
Marta Rodríguez (AENA)	Project Team	21/05/2021
Leila Zerrouki (EUROCONTROL)	Project Team	21/05/2021

Approved for submission to the SJU By - Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date
Rubén Alcolea (Nommon)	Project Coordinator	21/05/2021

Rejected By - Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date
-	-	-

Document History

Edition	Date	Status	Author	Justification
00.00.01	30/04/2021	Draft	Alex Gregg	Initial draft
00.01.00	21/05/2021	Submitted to the SJU for approval	Alex Gregg	Internal review before submission to the SJU

Copyright Statement

TRANSIT

TRAVEL INFORMATION MANAGEMENT FOR SEAMLESS INTERMODAL TRANSPORT

This deliverable is part of a project that has received funding from the SESAR Joint Undertaking under grant agreement No 893209 under European Union's Horizon 2020 research and innovation programme.

Abstract

This deliverable presents different methodologies and mobility analytics algorithms to reconstruct multimodal door-to-door passenger journeys from a variety of big data sources. The document starts by describing the different data sources that have been used for analysis. Then, we identify the main gaps in current approaches to passenger characterisation and propose a set of novel methodologies to solve such gaps. In particular, four main challenges are addressed: passenger segmentation by sociodemographic profile, analysis of modal choices in the airport access and egress legs, demand segmentation according to trip purpose, and characterisation of airport connectivity. The document concludes by discussing how the information extracted thanks to these new approaches will be used in the subsequent stages of the project to model long-distance, multimodal travel behaviour.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1. Scope and objectives

The goal of TRANSIT is to develop a set of multimodal key performance indicators (KPIs), mobility data analysis methods and transport simulation tools allowing the evaluation of the impact of innovative intermodal transport solutions on the quality, efficiency and resilience of the door-to-door passenger journey. The specific objectives of the project are the following:

1. Propose innovative intermodal transport solutions based on information sharing and coordinated decision-making between air transport and other transport modes.
2. Develop multimodal KPIs to evaluate the quality and efficiency of the door-to-door passenger journey.
3. Investigate new methods and algorithms for mobility data collection, fusion and analysis allowing a detailed reconstruction of the different stages of long-distance multimodal trips and the measurement of the new multimodal KPIs.
4. Develop a modelling and simulation framework for the analysis of long-distance travel behaviour that allows a comprehensive assessment of intermodal solutions in terms of the proposed multimodal KPIs.
5. Assess the expected impact of the proposed intermodal concepts and derive guidelines and recommendations for their practical development and implementation.

This document aims to answer the third of these objectives, looking at how to improve the characterisation of both demand and supply for long-distance travel.

In TRANSIT, mobile network data is proposed as the main data sources for the analysis of long-distance travel demand. Mobile network data is particularly suitable for this purpose thanks to the possibility of working with large, well-distributed population samples with high temporal and spatial resolution [1]. However, current approaches to the analysis of this data do not provide a number of key features about the profile of the users and the characteristics of the identified trips. The work described in the present document aims to fill some of these gaps:

- Mobile network data often lack reliable information about the sociodemographic characteristics of the users, such as age and gender. We propose to overcome this limitation by analysing the users' travel patterns and using machine learning techniques and other sources of data to estimate the users' age and gender. In subsequent project stages, this enhanced characterisation will be used as an input to model long-distance travel behaviour. Moreover, the proposed approach can be easily extended to other characteristics of the travellers that are correlated with their mobility patterns.
- Through the use of map-matching techniques, mobile network data can be used to identify the transport modes selected by the users in the different stages of a trip [1]. However, the applicability of this approach is limited by the similarity of the spatio-temporal trajectories of different transport modes: in urban and metropolitan areas, where different modes can generate very similar sequences of spatio-temporal registers, map matching techniques are far

less reliable. For airports whose location is close to an urban centre, this may make it difficult to analyse the modal share in the access and egress legs. To overcome this problem, we propose making use of survey data to calibrate a modal choice model that is then applied to the access legs identified from the mobile network data. This model helps identify the variables considered by the travellers in their mode choice decision and estimate their actual mode choice, enriching the information extracted from the mobile phone data.

- Finally, we investigate how to use the mobile network data to infer long-distance travel purposes. By analysing longitudinal mobile network data time series and combining them with airport surveys, we will calibrate a machine learning model to identify trip purpose in long-distance trips, classifying them into business and leisure.

These improvements will be put at work in the next project stages to reconstruct the travel demand information required to build and calibrate the TRANSIT modelling framework.

Regarding the characterisation of transport supply, we also investigate how to use historical flight data to characterise airport connectivity and resilience, with the aim to use this information in subsequent work when evaluating and assessing different intermodal concepts.

1.2. References

[1] García, P., Herranz, R., & Javier, J. (2016). Big data analytics for a passenger-centric air traffic management system. 6th SESAR Innovation Days, Delft, Netherlands.

[2] Horl, S., & Balac, M. (2020). Open data travel demand synthesis for agent-based transport simulation: A case study of Paris and Île-de-France. https://doi.org/10.3929/ethz-b-000412979

[3] Namazi-Rad, M. R., Tanton, R., Steel, D., Mokhtarian, P., & Das, S. (2017). An unconstrained statistical matching algorithm for combining individual and household level geo-specific census and survey data. Computers, Environment and Urban Systems, 63, 3-14.

[4] Ye, X., Konduri, K., Pendyala, R. M., Sana, B., & Waddell, P. (2009, January). A methodology to match distributions of both household and person attributes in the generation of synthetic populations. In 88th Annual Meeting of the Transportation Research Board, Washington, DC..

[5] Sánchez, O., Isabel, M., & González, E. M. (2014). Travel patterns, regarding different activities: work, studies, household responsibilities and leisure. Transportation Research Procedia, 3, 119-128.

[6] Hwang, H., Wilson, D., Taylor, R., & Chin, S. (2015). Travel patterns and characteristics of the elderly subpopulation in New York state. Oak Ridge National Laboratory, US.

[7] Tilley, S., & Houston, D. (2016). The gender turnaround: Young women now travelling more than young men. Journal of transport geography, 54, 349-358.

[8] Ng, W. S., & Acker, A. (2018). Understanding urban travel behaviour by gender for efficient and equitable transport policies. International Transport Forum Discussion Paper.

[9] Lenormand, M., Louail, T., Cantú-Ros, O. G., Picornell, M., Herranz, R., Murillo, J., Barthelemy, M., San Miguel, M. & Ramasco, J. J. (2015). Influence of sociodemographic characteristics on human mobility. Scientific reports, 5(1), 1-15.

[10] Gauvin, L., Tizzoni, M., Piaggesi, S., Young, A., Adler, N., Verhulst, S., Ferres, L., & Cattuto, C. (2020). Gender gaps in urban mobility. Humanities and Social Sciences Communications, 7(1), 1-13.

[11] Herrera-Yagüe, C., & Zufiria, P. J. (2012). Prediction of telephone user attributes based on network neighborhood information. In International Workshop on Machine Learning and Data Mining in Pattern Recognition (pp. 645-659). Springer, Berlin, Heidelberg.

[12] de Dios Ortúzar, J., & Willumsen, L. G. (2011). Modelling transport. John Wiley & Sons.

[13] Sarraute, C., Blanc, P., & Burroni, J. (2014, August). A study of age and gender seen through mobile phone usage patterns in Mexico. In 2014 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM 2014) (pp. 836-843). IEEE.

[14] O. A. Worldwide, Schedule Analyzer | New Airline Routes Planner | OAG. https://www.oag.com/schedules-analyzer (accessed 22nd April 2021).

[15] IATA (2020). Air Connectivity.

[16] Veldhuis, J. (1997). The competitive position of airline networks. Journal of air transport management, 3(4), 181-188..

[17] Burghouwt, G., & de Wit, J. (2005). Temporal configurations of European airline networks. Journal of Air Transport Management, 11(3), 185-198..

[18] Doganis, R., & Dennis, N. (1989). Lessons in hubbing. Airline Business..

[19] Bootsma, P. D. (1997). Airline Flight Schedule Development-Analysis and design tools for European hinterland hubs.

[20] Danesi, A. (2006). Measuring airline hub timetable co-ordination and connectivity: definition of a new index and application to a sample of European hubs.

[21] Assemblée Nationale, Projet de loi no 3875 portant lutte contre le dérèglement climatique et renforcement de la résilience face à ses effets », Assemblée nationale. www.assemblee-nationale.fr/dyn/15/textes/l15b3875_projet-loi (accessed 23rd April 2021).

[22] CDG Facile, Guide Correspondance à CDG, Transit », Aéroport Roissy Charles de Gaulle (Paris-CDG), 2021. https://cdgfacile.com/informations-passagers-correspondance-cdg/ (accessed 23rd April 2021).

[23] Commissariat général à la stratégie et à la perspective (2013). Valeur du temps.

[24] Jacquillat, A., & Odoni, A. R. (2015). An integrated scheduling and operations approach to airport congestion mitigation. Operations Research, 63(6), 1390-1410.

[25] Reiners, T., Pahl, J., Maroszek, M., & Rettig, C. (2012, January). Integrated aircraft scheduling problem: An auto-adapting algorithm to find robust aircraft assignments for large flight plans. In 2012 45th Hawaii International Conference on System Sciences (pp. 1267-1276). IEEE.

[26] Air Journal. Gilets jaunes : baisse des arrivées de touristes à Orly et Roissy-CDG. https://www.air-journal.fr/2019-01-12-gilets-jaunes-baisse-des-arrivees-de-touristes-a-orly-et-roissy-cdg-5209613.html (accessed 22nd April 2021).

[27] Schafer, R. W. (2011). What is a Savitzky-Golay filter? [Lecture notes]. IEEE Signal processing magazine, 28(4), 111-117.

1.3. List of acronyms

Table 1.1. List of acronyms

Acronym	Definition
ATC	Air Traffic Control
CDG	Charles de Gaulle
DBSCAN	Density-based spatial clustering of applications with noise
GB	Gradient Boosting
GTFS	General Transit Feed Specification
HST	High Speed Trains
IATA	International Air Transport Association
IPF	Iterative Proportional Fitting
IPU	Iterative Proportional Updating
IVT	In-Vehicle Time
KNN	k-Nearest Neighbours
KPA	Key Performance Areas
KPI	Key Performance Indicators
ML	Machine learning
MLP	Multilayer Perceptron
OD	Origin-Destination
OSRM	Open-Source Routing Machine
OPT	Open Trip Planner
PCA	Principal Component Analysis
PT	Public Transport
RF	Random Forest
RDPS	Radar Data Processing System
SES	Single European Sky
SESAR	Single European Sky ATM Research
SMOTE	Synthetic Minority Oversampling Technique
SNCF	Société Nationale des Chemins de Fer français
SVM	Support Vector Machines
TTP	Tarjeta de Transporte Público (Public Transport Travel Card)
WP	Work Package

1.4. Document structure

The rest of this document is structured as follows:

- **Section 2** presents the main data sources used for the different analyses.
- **Section 3, Section 4 and Section 5** describe the methodologies developed to infer travellers' sociodemographic profile, airport access modal choices, and long-distance travel purposes. For each development, we formalise the problem, describe the proposed validation plan and present the main results obtained for the case study of the Madrid-Barajas airport.
- **Section 6** describes the methodology developed to evaluate airport connectivity and aircraft flows and presents the results obtained for the case study of Paris-Charles-de-Gaulle airport.
- **Section 7** summarises the main conclusions of the study and discusses how these results will be used in the next work packages of the TRANSIT project.

2. Data sources

The main data sources used in the present study are the following:

- Mobile network data obtained through a collaboration agreement with Orange Spain.
- Smart card data from the automated fare collection system of the Madrid transport network.
- AENA passenger surveys (EMMA surveys)
- Madrid Household Travel Survey (EDM2018)
- OAG historical flight data
- RDPS radar data.

The characteristics of these data sources (access conditions, content, limitations, spatial and temporal scope, resolution, etc.) are described in detail in the data factsheets included in Appendix A.

3. Enrichment of passenger profiling

3.1. Problem statement

One of the main objectives of TRANSIT is to develop a modelling framework that allows the simulation of the impact of different intermodal concepts and information services on multimodal long-distance travel behaviour. This simulation model will require disaggregated data to an agent level, which implies that a synthetic population reproducing the main sociodemographic characteristics of the real population needs to be created.

As introduced in the previous section, mobile network data has become a fundamental tool to extract the mobility patterns of the population and reconstruct the activity-travel diaries of the users. Although mobile phone data contains information on the age and gender of some users, many users do not have any sociodemographic values attached to them or their values are incorrect (e.g., young people whose contract data correspond to one of their parents). This kind of information is of paramount importance for the development of meaningful and actionable information about people's mobility and for the development of modal choice models. The lack of this kind of values makes it necessary to develop approaches able to estimate these characteristics in a reliable manner.

To this end, a machine learning model has been developed that uses as input variables the mobility patterns of the population in order to estimate the and gender of the travellers. The approach presented in this document differs from the techniques usually covered in the literature for generating synthetic populations, which normally consist of statistical matching techniques ([2],[3]) and iterative approaches like IPF and IPU ([4]).

The rest of this chapter is organised as follows: Section 3.2 describes the methodology applied to develop the model and the main insights obtained about the relationship between mobility indicators and the sociodemographic characteristics of the travellers. Section 3.3 presents the validation plan. Section 3.4 describes the results of the final predictive model. Section 3.5 shows an application of the model to the passengers travelling to and from Madrid-Barajas airport. Finally, Section 3.6 presents the main conclusions on the new profiling method.

3.2. Methodology

From data exploration to the calibration of the final machine learning model, this section covers all the experiments conducted in developing the solution. The main steps followed are:

1. Select the input variables of the problem. From the beginning, the focus was set on using the mobility patterns of the population as predictors of their age and gender. To that end, a literature review was conducted to select those indicators where there are more differences between the segmentations of both variables. More concretely, gender was divided into male and female users, and age was segmented into 10 groups comprising 10 years each. Once the literature variables were identified, a survey analysis was conducted to select those patterns which are more characteristic of the difference between the variables.
2. Cluster the individuals of the surveys in terms of their mobility patterns, so as to divide the individuals into *k* groups and analyse the age and gender distributions of each group to see if

the clustering technique was able to divide correctly the survey participants and, in that case, to develop a set of rules to classify each individual into the different age and gender groups. As it will be explained later, although the distributions are not as separable as expected, the experiment was of great use to identify new patterns that were used in later steps.

3. Cluster the users of the mobile phone dataset. Once the participants of the surveys were clustered, the same procedure was applied to the users present in the mobile phone data sample, to see if a more separable division could be obtained. The results were similar to the ones obtained from the surveys.
4. Development of supervised machine learning models. Once it was clear that it was not going to be possible to build a set of rules from the clustering results that could separate accurately the users in the different age and gender groups, the perspective of the problem changed and a supervised approach was sought. Two machine learning models were developed, one capable of predicting the age of the mobile phone data sample and another capable of predicting gender. For both models, the variables used as predictors were the mobility patterns selected in step 1. To calibrate both models, the classical steps of the machine learning life cycle were followed: data preparation, data wrangling, data analysis, model training with hyperparameter tuning, and model testing.

3.2.1. Analysis of survey data to extract mobility patterns

There are a number of individual mobility variables that the literature suggests that present important variations across different age and gender groups. Some of these variables can be also observed for the agents sampled from mobile network data. This is an opportunity for refining the age and gender information provided for the mobile phone users. The objective of this first step of the algorithm was to analyse the available surveys to identify if the differences mentioned in the literature also hold for the survey participants. The surveys available for this study were the household surveys from Madrid (2018), Valencia (2019) and Seville (2017).

The main individual mobility indicators obtained from the literature were: number of trips, trip purpose, trip distance, trip duration, radius of gyration, and the temporal distribution of the trips. These patterns have been explored for the participants of the three surveys mentioned above, and used to see if they imply any differences in age and/or gender.

Trips per person

The average number of trips per person is usually higher for women than for men, as shown in [4] and [5]. Although the difference is not significant, we can see this pattern in the three surveys. Concretely, in Madrid's survey the value is 4% higher for women, while in Valencia and Seville it is 3% and 1%, respectively. Regarding age, the differences are small for the three surveys.

Figure 3.1 shows the trips per person distribution in Madrid's survey for all the age and gender groups considered. A recurrent pattern is that the eldest population groups show a smaller number of daily trips. Despite the fact that the variability is limited, the pattern of trips per person was selected for the clustering step, as the mobile phone data sample from Orange is known to have more variability, probably as a result of the higher sample size and the fact that household surveys are revealed preference sources sometimes suffering from underreporting of trips.

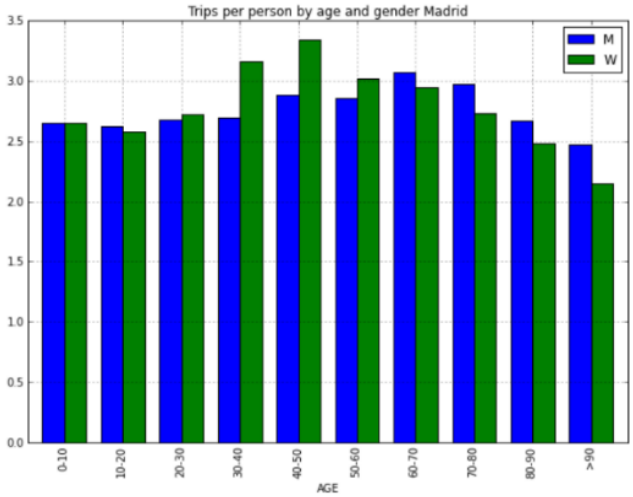


Figure 3.1: Trips per person distribution per age and gender for Madrid

Trip distance

Trip distance was the indicator mentioned with the most recurrence in the literature. In [4] it was found that the average trip distance for men was 37% higher than for women in Andalusia. Also, the indicator suffers a significant decrease after reaching the age of 60. Other papers that mention similar patterns are [6], [7] and [8]. After studying the survey data, it was confirmed as the most differentiating pattern between different groups of age and gender. In the survey of Madrid, the average distance per trip is 18% higher for men than for women, while in Valencia and Seville the difference is 20% and 7% higher for men, respectively.

As it can be seen in Figure 3.2 for the survey of Madrid, the biggest differences in gender appear in the active population groups (mostly from 30 to 60). Regarding age, the average trip distance decreases significantly outside these groups (0-20 and >60). The distance distribution for Valencia and Seville was found to be very similar to the one for Madrid.

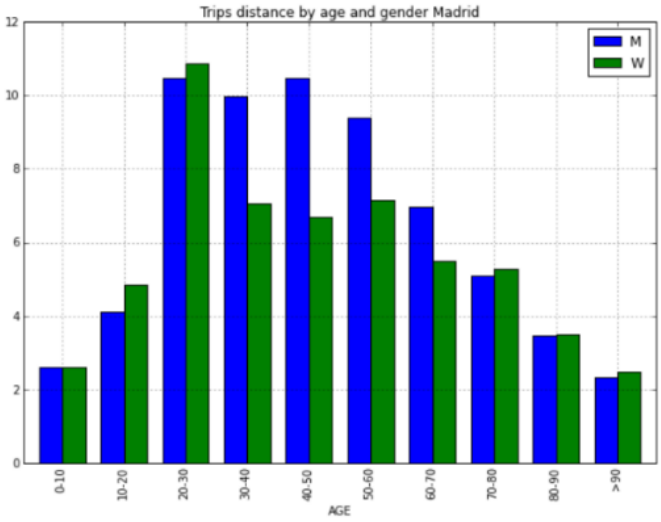


Figure 3.2: Trip distance distribution per age and gender for Madrid

Trip purpose

In the study conducted in [4], it was identified that men make 20% more trips than women both for business and leisure activities, while women make more trips related to household responsibilities and studying. However, these differences were not identified in the surveys. The distribution per gender by trip purpose for Madrid's survey is shown in Figure 3.3, where one can see that there are not any significant changes between different types of activities. The figure includes the following purposes:

- H: home
- W: work, studies
- L1: shopping
- L2: accompanying someone
- L3: leisure activities
- L4: sport activities
- L5: personal affairs
- O: doctor, others.

The same thing happens for the different ranges of age that have been considered.

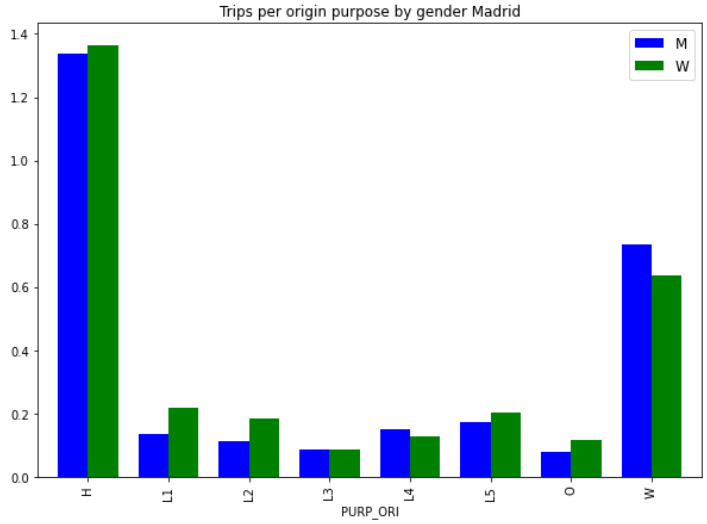


Figure 3.3: Trip purpose distribution by gender for Madrid

Trip duration

Trip duration tends to be correlated with trip distance in most of the individuals of the population, except from elderly people. Therefore, regarding gender, one would expect the distribution of trip duration by gender to be similar to the one obtained for trip distance. As it can be seen in Figure 3.4, which shows the trip distance distribution for Madrid, this has been confirmed. However, differences between men and women of the active population tend to me smaller in trip duration than trip distance.

Regarding age, there is an increase in travel time compared to trip distance in the groups comprising elderly people. However, the differences are not as high as the ones seen with trip distance, and

therefore this indicator was not selected for the clustering step. In addition, the measurement of trip durations through mobile network data is not as accurate as the measurement of trip distances, due to the uncertainty about the start and end time of the activities in relation to the registers (the activity at the origin may have finished some minutes after the last register in the activity location, and vice versa for destination).

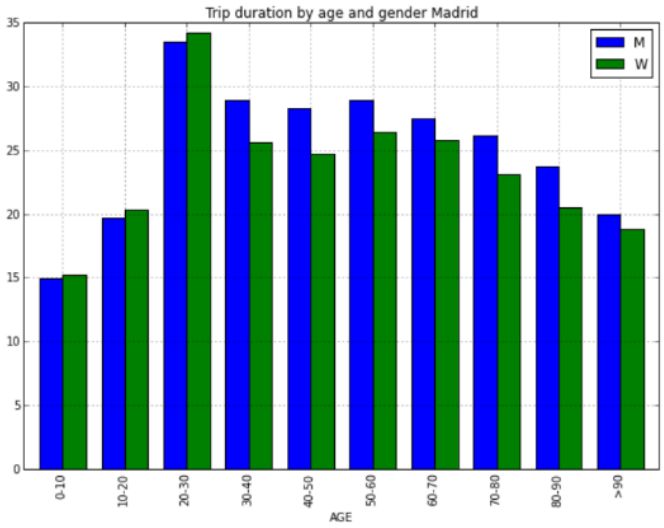


Figure 3.4: Trip duration distribution per age and gender for Madrid

Radius of gyration

The radius of gyration can be defined as:

$$r_g = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(\overrightarrow{p_k} - \overrightarrow{p_c})^2},$$

where $\overrightarrow{p_k}$ represents the k^{th}position of the individual displacements and $\overrightarrow{p_c} = \frac{1}{n}\sum_{k=1}^{n}\overrightarrow{p_k}$ is the centre of mass of his/her motions.

The tendency followed by this indicator is also similar to the one observed for trip distance. In the literature, the main patterns identified include that women trajectories stay closer to their centre of mass and the fact that the radius of gyration decreases with age ([8],[9]). These patterns were confirmed after studying the surveys, as the radius of gyration was 13%, 12% and 5% higher for men for the Madrid, Valencia and Seville surveys, respectively, and the indicator experienced a significant decrease from the active population group to the elderly ones. This can be seen in Figure 3.5. As the differences are more significant than those of trip duration, this indicator was selected for the clustering step.

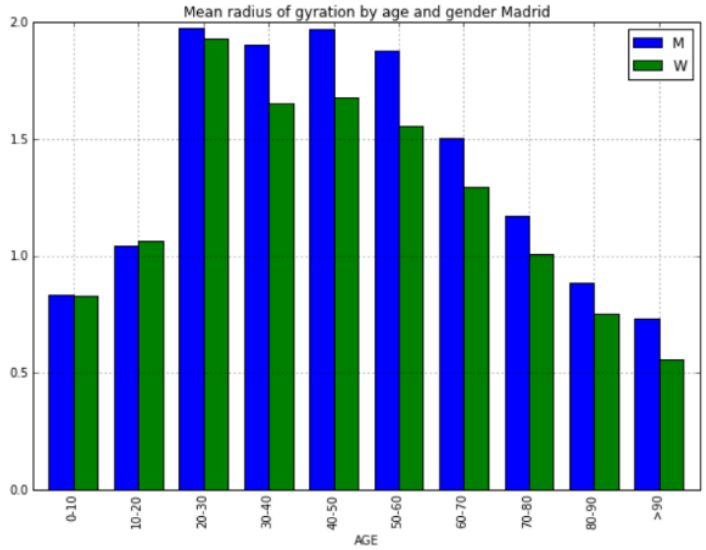


Figure 3.5: Radius of gyration distribution per age and gender for Madrid

Temporal distribution

The last variable analysed refers to the trip distribution by time of the day. After studying the surveys, it was seen that women take a higher number of trips in off-peak hours, which can be interesting for our model. For the age distribution, which can be seen in Figure 3.6, three main age groups with similar tendencies can be easily identified: young population (from 0 to 20 years), which have the majority of their trips in the morning peak, active population (from 30 to 60 years), which have a distribution compatible with a home-work-home pattern, and elderly population (from 60 to >90 years), which have most of their trips in off-peak hours. Due to the differences observed between these population groups, this variable was selected for the clustering procedure.

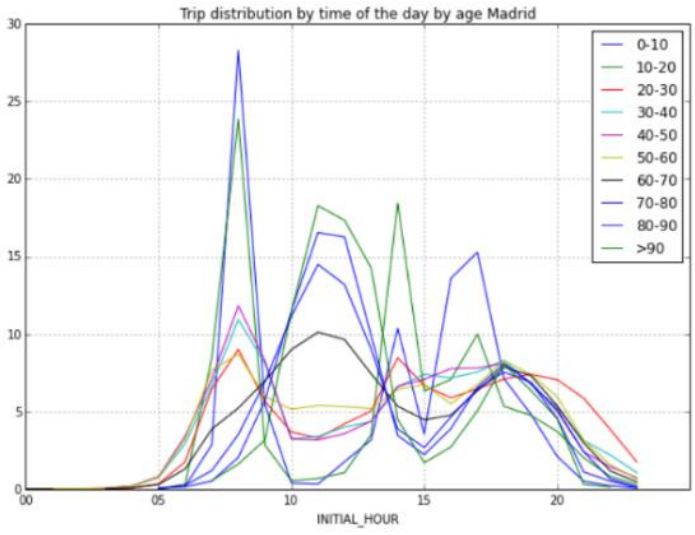


Figure 3.6: Trip distribution by time of the day for different age groups in Madrid

3.2.2. Clustering application on survey data

Based on the previous analysis, the following indicators were selected for the application of the clustering procedure: number of trips, trip distance, radius of gyration and temporal distribution.

There are several clustering techniques that can be applied to our problem, such as k-means, hierarchical clustering and local methods like DBSCAN. Given the size of the surveys (more than 70K samples for Madrid, 15K for Valencia and 4K for Seville), the most efficient method is k-means.

The algorithm behind k-means clustering requires indicating the number of clusters *k* to be created. Another condition is that all variables of each sample must be numeric, as it has to compute distances between individuals of the sample. The method starts by randomly choosing a set of centroids for each cluster, and assigning each sample to the cluster whose centroid is more similar to it. Then the new centroids are calculated as the mean of the points in each cluster. The process continues iteratively until the centroids are sufficiently stabilised. To obtain the optimal number of clusters, the algorithm is run for different values of *k*, and for each one of the results, a score is obtained to measure the quality of the clusters. There are several score metrics that can be used for this kind of purpose, being the most typical and the one selected in this experiment the *silhouette score*, which measures in average how close each sample is to its cluster and how far it is from the rest of the clusters, returning a number between -1 and 1. When the silhouette score is close to 1, it means that the data is well distributed in their clusters. A silhouette value near -1 implies that in general the samples have been distributed to the wrong cluster, while a silhouette score of 0 means that the sample is in the border between two clusters.

The experiment was conducted separately for each of the surveys. From the set of mobility indicators selected, the only one that is not clearly represented as a numeric value is the temporal distribution of the trips. In order not to use too many variables, this pattern was divided into two categories: “number of peak-hour trips” and “number of off-peak hour trips”. Therefore, the final variables used in the clustering method for each user in the surveys were: number of trips in the survey, average trip distance, radius of gyration, number of peak-hour trips and number of off-peak hour trips.

For each survey, the analysis was divided into the following steps:

1. Extracting the mobility patterns for each individual in the survey.
2. Pre-processing the variables selected to input them to the k-means algorithm.
3. Applying k-means and extracting the optimal number of clusters *k*.
4. Analysing the results in terms of the age and gender distribution in each one of the clusters selected for the optimal value of *k*.

Extraction of mobility patterns

This was the easiest step of the process, as most of the patterns had already been calculated in the previous analysis. The only new pattern that needed to be calculated at this point was the number of “peak” and “off-peak” trips. The hour division selected to obtain these variables was:

```
"periods": {
        “peak”: ["06", "07", "08", "09", "14", "15", "18", "19"],
        “off_peak”: ["00", "01", "02", "03", "04", "05", "10", "11", "12", "13", "16", "17", "20", "21", "22", "23"]
 }
```

Variable pre-processing

After obtaining the five variables mentioned above, their distributions were analysed. Some particularly extreme outliers were found for Madrid's survey, with distance values above 200 km. As these distances were exceptionally high in comparison to the rest of samples of the survey, they were removed from the dataset. After that, the distribution of all the variables was measured, to see if there were any more outliers. In the following shown in Figure 3.7, it can be seen that there are still many outliers for all the variables.

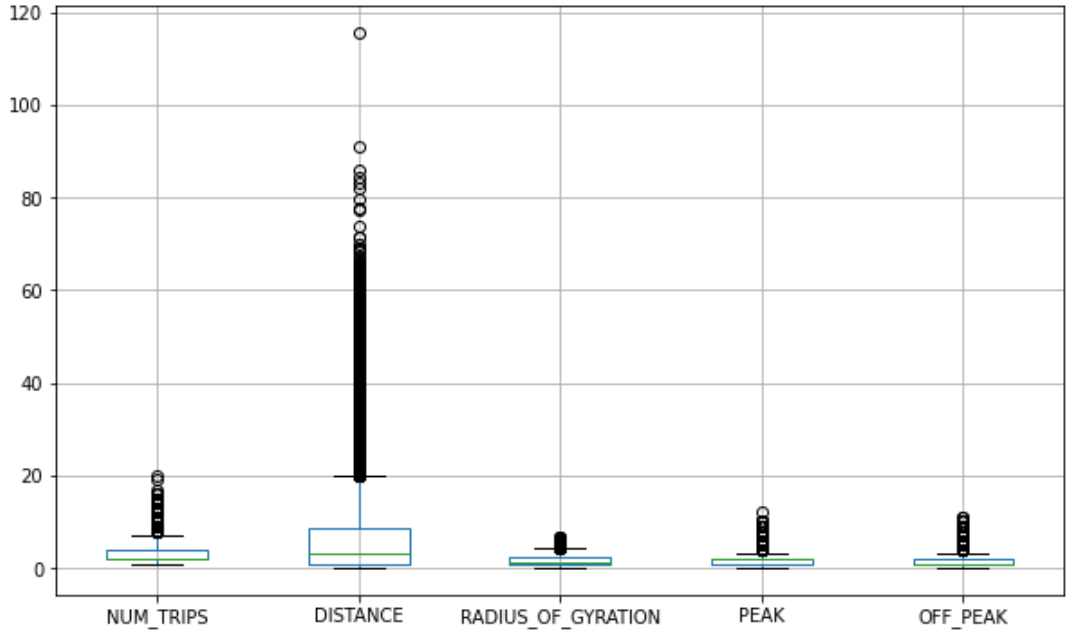


Figure 3.7: Distribution of clustering variables

As a first approach, the outliers were kept for the clustering step, but the results showed a very poor silhouette score (~0.2) on all possible values of the number of clusters *k*. Therefore, from this point, and for all the surveys, the outliers were removed and the final sample size in each case was 60K for Madrid, 13K for Valencia and 3K for Seville.

Another aspect to take into account is that the K-Means algorithm works by computing distances, and this can cause problems if the input variables are not all in the same scale. For example, as the average distances are higher than the number of trips, the former would have more impact on deciding the centroid of the cluster that is the "closest" to each sample. To normalize the variables, the approach followed was the standardization, which consists in two steps: 1) calculate the mean and standard deviation for the variable and 2) for each observed value of the variable, subtract the mean from it and divide the subtraction by the standard deviation. The process generates standard values that represent the number of standard deviations by which the value is above or below the mean of a particular variable.

K-means clustering

Once we had pre-processed the variables, we were ready to use them for the clustering step. For the implementation of the algorithm, the Python library used was *scikit-learn*, and its sub-module *cluster*, which includes the function named *KMeans*. This function receives as parameters the number of clusters *k*, the maximum number of iterations to run (set by default to 300) and the random seed to generate the initial random centroids (set to 0 by default). The number of clusters k tested in each survey ranged from 2 to 10. Table 3.2 shows the values obtained for the silhouette scores of each survey for each value of *k*.

Table 3.2: Silhouette score values obtained for each survey

	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10
Madrid	0.29	0.35	0.35	0.33	0.34	0.36	0.36	0.38	0.39
Valencia	0.29	0.36	0.38	0.39	0.39	0.38	0.41	0.41	0.43
Seville	0.40	0.40	0.41	0.48	0.50	0.54	0.51	0.54	0.57

It appears that the higher the value of *k*, the higher the silhouette score. On the other hand, it can be seen that the scores are not sufficiently high, so the clustering technique is not being able to separate the clusters enough. To solve this problem, a reduction of the number of input was performed, keeping only the ones that made a more significant impact when it comes to differentiating the samples. The variables removed were the number of "peak" and "off-peak" trips, keeping therefore the number of trips, the average distance and the radius of gyration of each individual. After applying this modification, the k-means algorithm was run again for each survey. Table 3.3 shows the results obtained for each value of *k*:

Table 3.3: Silhouette score values for each survey after removing "peak" and "off-peak" variables

	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10
Madrid	0.44	0.46	0.43	0.46	0.46	0.43	0.44	0.45	0.44
Valencia	0.45	0.52	0.49	0.52	0.53	0.52	0.53	0.55	0.54
Seville	0.57	0.53	0.57	0.60	0.59	0.58	0.58	0.59	0.58

In this second implementation, it can be seen that the silhouette scores for each survey have been increased, being Seville's survey the case where the maximum value is reached (0.60 for k=5). In addition, for all the surveys, except for small differences in the case of Valencia, the best values for the silhouette score are concentrated in the middle values of *k.*

The last step of this clustering process consists in selecting the optimal value of *k* for each survey. With the information extracted from Table 3.2, the value selected for all surveys was *k*=5. In addition to being the value of *k* which provides the best results (except for Valencia), having the same value for every survey will facilitate the posterior analysis of the results.

Analysis of results

The objective of the whole clustering analysis was to investigate the age and gender distributions of the clusters generated and see if the algorithm had been able to segment the individuals of different age groups into different clusters.

To be able to see the differences more clearly, the approach followed consisted in plotting both the age and gender distributions together per each age and gender class. Figure 3.8, Figure 3.9 and Figure 3.10 show the distributions for each survey. As it can be seen, the values of the ages comprising 30 to 60 are almost always higher than the rest, as the surveys mostly contain people between those ages.

There are some important differences between the clusters that must be highlighted. In Madrid's survey, most elderly people and a high percentage of the youngest individuals are represented in cluster 5. This indicates that there might be young users that behave as elderly people when it comes to mobility. This is something that also happens in Valencia (cluster 1) and Seville (cluster 1). This suggests that this may be an interesting pattern to be analysed further. The rest of the clusters are mostly dominated by the groups mentioned above, with higher or lower presence from young and elderly people.

Further analysis on the dissimilarities between different age groups is shown in the following section. Changing the focus to gender, it can be seen that in all the surveys the number of women that participated is higher than the number of men. However, there are some clusters where men are more represented, like cluster 1 in Valencia. Unfortunately, there are not enough differences between clusters with respect to gender. This can be seen in Figure 3.11, which shows that the distribution of men and women in each of the five clusters obtained from Madrid's survey is highly similar between them.

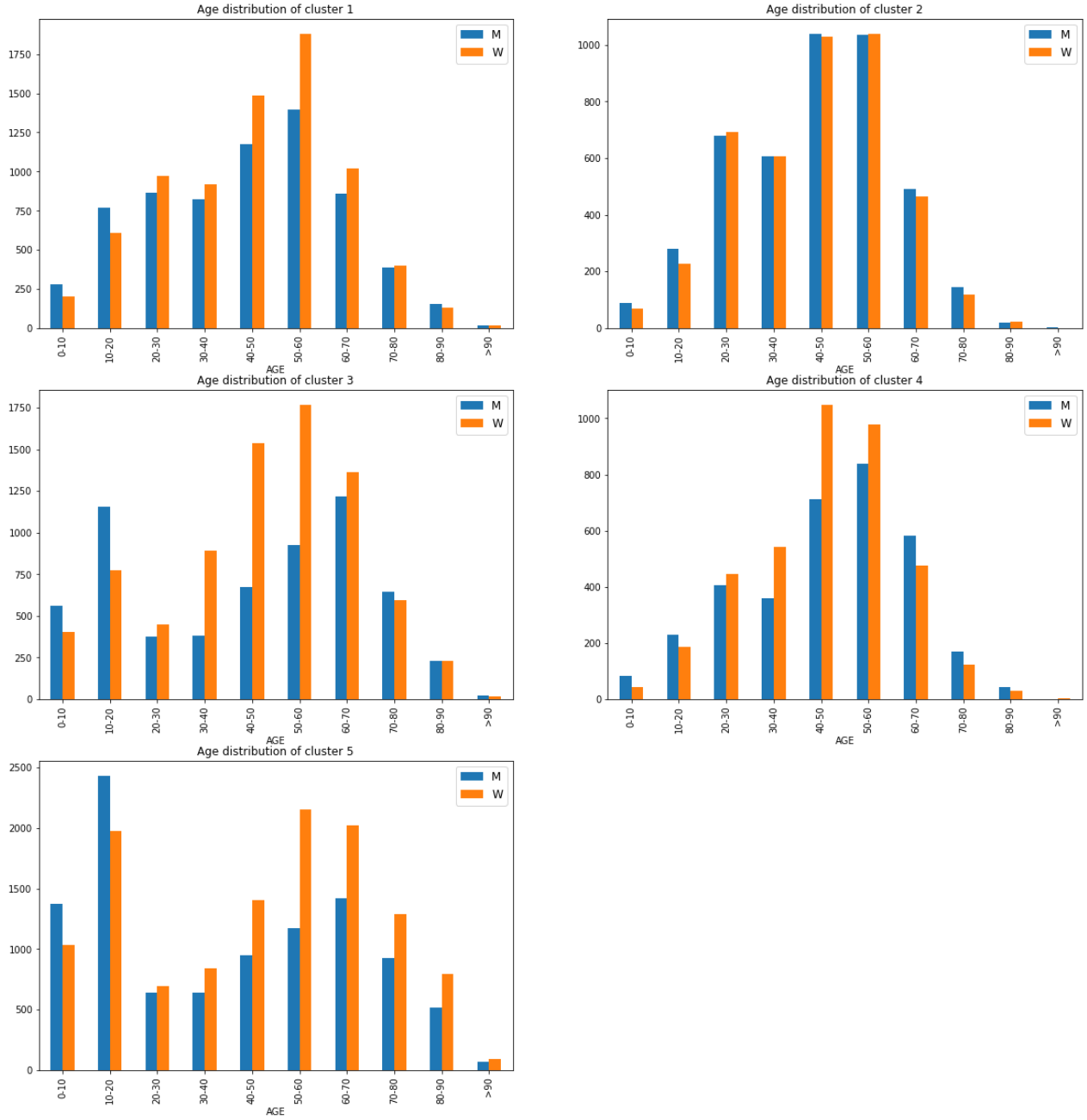


Figure 3.8: Distribution of age and gender per cluster in Madrid's survey

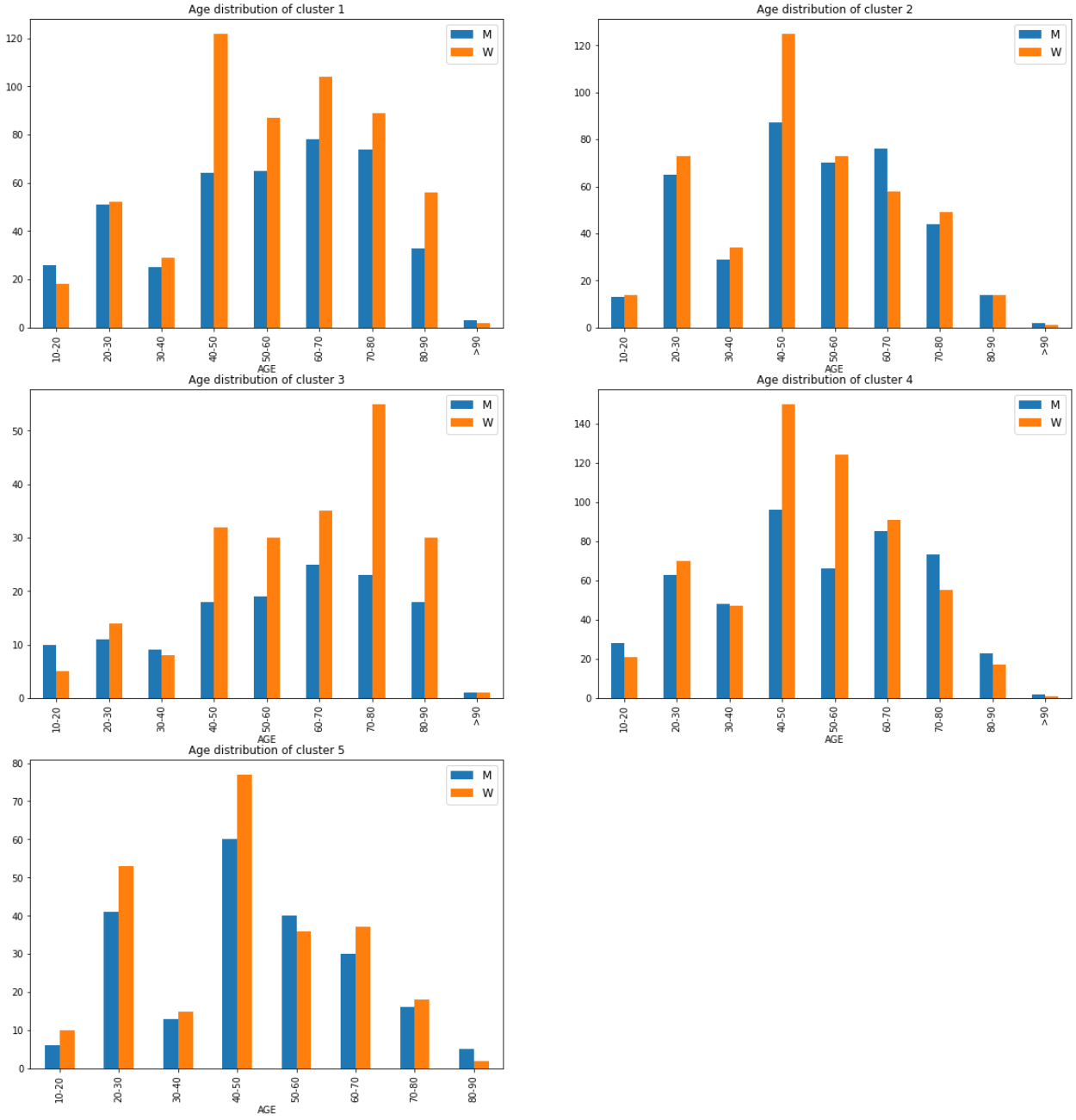


Figure 3.9: Distribution of age and gender per cluster in Valencia's survey

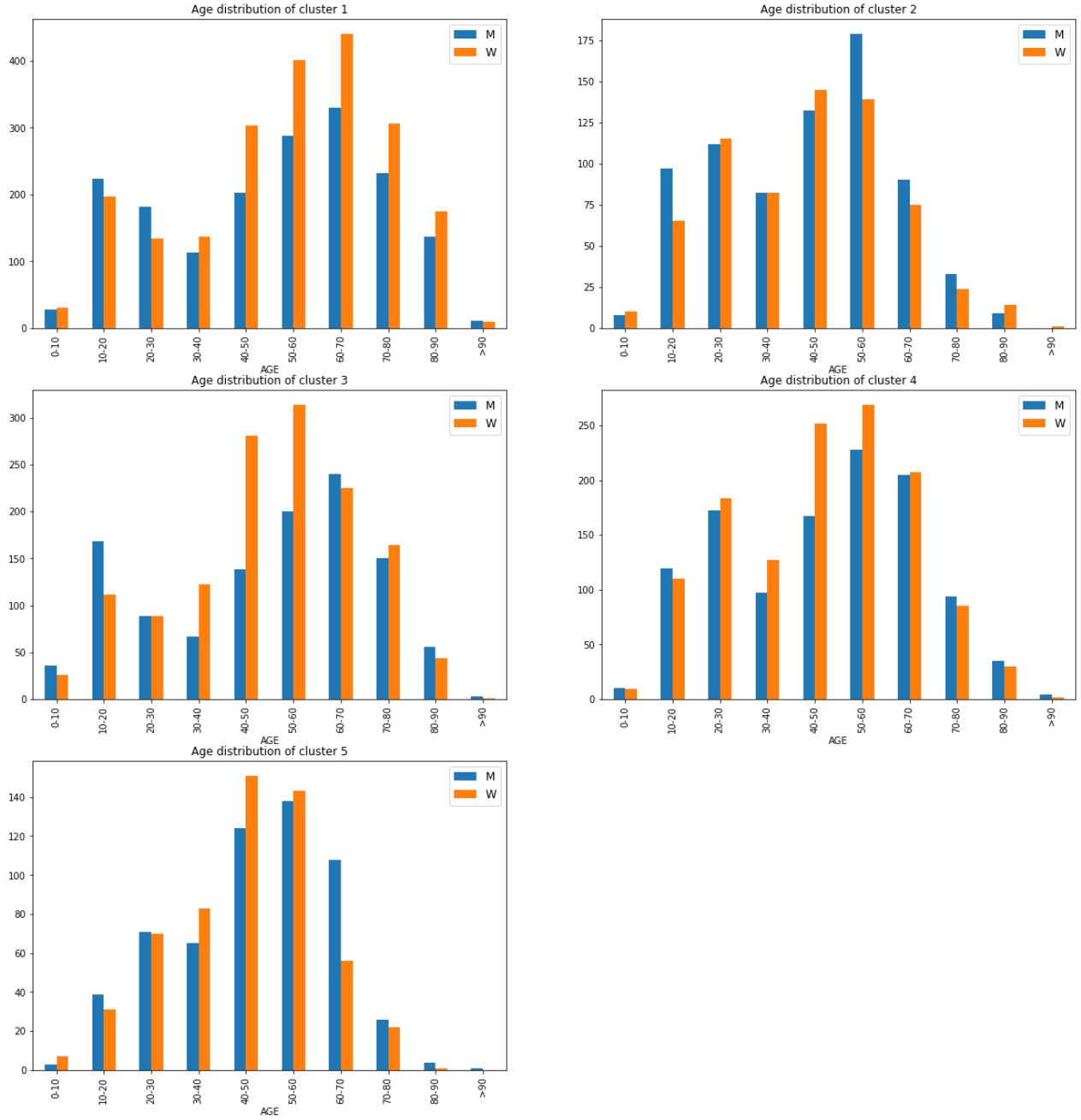


Figure 3.10: Distribution of age and gender per cluster in Seville's survey

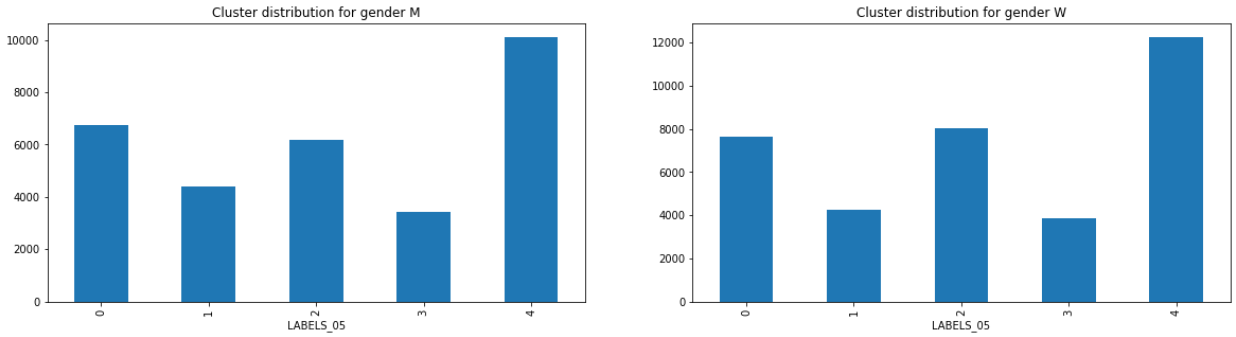


Figure 3.11: Distribution by cluster for men and women in Madrid's survey

Comparison of age clusters across cities

Once the results of the clustering process were examined separately for each survey, the age clusters obtained in the three analysis were compared together in order to see to what extent they were similar and if this could suggest further tests with other variables.

Table 3.4, Table 3.5 and Table 3.6 show the distribution of respondents' age groups across the five clusters obtained from the household surveys. It has to be noted that some structural similarities can be found, in particular the existence of a cluster joining the youngest and eldest age ranges (cluster 5 in Madrid, and cluster 1 in Seville and Valencia).

Table 3.4: Distribution of age groups across clusters in Madrid's survey

MAD	CLUSTER_1_MAD	CLUSTER_2_MAD	CLUSTER_3_MAD	CLUSTER_4_MAD	CLUSTER_5_MAD	
0-10	11,7%	3,8%	23,3%	3,1%	58,1%	100,0%
10-20	15,9%	5,9%	22,4%	4,8%	51,0%	100,0%
20-30	29,6%	22,1%	13,3%	13,7%	21,4%	100,0%
30-40	26,4%	18,4%	19,2%	13,6%	22,3%	100,0%
40-50	24,1%	18,7%	20,0%	15,9%	21,3%	100,0%
50-60	24,9%	15,7%	20,4%	13,8%	25,2%	100,0%
60-70	18,9%	9,7%	26,1%	10,6%	34,7%	100,0%
70-80	16,4%	5,5%	25,8%	6,1%	46,2%	100,0%
80-90	13,2%	1,9%	21,1%	3,2%	60,5%	100,0%
>90	14,3%	0,9%	16,1%	1,3%	67,4%	100,0%
	21,5%	12,9%	21,2%	10,9%	33,5%	n=66914

Table 3.5: Distribution of age groups across clusters in Seville's survey

SVQ	CLUSTER_1_SVQ	CLUSTER_2_SVQ	CLUSTER_3_SVQ	CLUSTER_4_SVQ	CLUSTER_5_SVQ	
10-20	29,1%	17,9%	9,9%	32,5%	10,6%	100,0%
20-30	20,9%	28,0%	5,1%	27,0%	19,1%	100,0%
30-40	21,0%	24,5%	6,6%	37,0%	10,9%	100,0%
40-50	22,4%	25,5%	6,0%	29,6%	16,5%	100,0%
50-60	24,9%	23,4%	8,0%	31,1%	12,5%	100,0%
60-70	29,4%	21,6%	9,7%	28,4%	10,8%	100,0%
70-80	32,9%	18,8%	15,7%	25,8%	6,9%	100,0%
80-90	42,0%	13,2%	22,6%	18,9%	3,3%	100,0%
>90	38,5%	23,1%	15,4%	23,1%	0,0%	100,0%
	26,6%	22,8%	9,3%	28,8%	12,5%	n=3862

Table 3.6: Distribution of age groups across clusters in Valencia's survey

VLC	CLUSTER_1_VLC	CLUSTER_2_VLC	CLUSTER_3_VLC	CLUSTER_4_VLC	CLUSTER_5_VLC	
0-10	28,3%	8,8%	30,2%	9,3%	23,4%	100,0%
10-20	33,4%	12,9%	22,2%	18,2%	13,2%	100,0%
20-30	27,8%	20,0%	15,7%	31,2%	5,4%	100,0%
30-40	26,7%	17,5%	20,1%	23,9%	11,8%	100,0%
40-50	27,5%	15,1%	22,8%	22,8%	11,8%	100,0%
50-60	29,7%	13,7%	22,1%	21,4%	13,0%	100,0%
60-70	35,4%	7,6%	21,4%	18,9%	16,7%	100,0%
70-80	38,9%	4,1%	22,7%	12,9%	21,4%	100,0%
80-90	46,3%	3,4%	14,9%	9,7%	25,7%	100,0%
>90	39,6%	1,9%	7,5%	11,3%	39,6%	100,0%
	32,4%	11,8%	21,1%	20,1%	14,7%	n=11981

To check the similarity between the three clustering results, the fifteen clusters (five per city) were compared by correlating the proportion of users from each age group that falls within the clusters (correlation between the values in each column in the above tables).

Table 3.7 shows the correlation coefficients obtained. As can be seen, there are many coefficients above 0.8 or below -0.8, indicating important similarities.

Table 3.7: Correlation between the clusters age structures across cities

	CLUSTER_1_SVQ	CLUSTER_2_SVQ	CLUSTER_3_SVQ	CLUSTER_4_SVQ	CLUSTER_5_SVQ
CLUSTER_1_MAD	-0,923	0,836	-0,858	0,590	0,849
CLUSTER_2_MAD	-0,943	0,795	-0,870	0,603	**0,906**
CLUSTER_3_MAD	0,246	-0,575	0,288	0,057	-0,209
CLUSTER_4_MAD	-0,907	0,716	-0,833	0,614	0,874
CLUSTER_5_MAD	0,939	-0,711	0,853	-0,660	-0,896

	CLUSTER_1_VLC	CLUSTER_2_VLC	CLUSTER_3_VLC	CLUSTER_4_VLC	CLUSTER_5_VLC
CLUSTER_1_MAD	-0,655	0,848	-0,055	0,956	-0,770
CLUSTER_2_MAD	-0,759	0,904	0,129	0,943	**-0,841**
CLUSTER_3_MAD	0,250	-0,430	0,624	-0,494	0,082
CLUSTER_4_MAD	-0,681	0,795	0,199	0,866	-0,811
CLUSTER_5_MAD	0,708	-0,833	-0,236	-0,895	0,861

	CLUSTER_1_SVQ	CLUSTER_2_SVQ	CLUSTER_3_SVQ	CLUSTER_4_SVQ	CLUSTER_5_SVQ
CLUSTER_1_VLC	0,981	-0,829	0,975	-0,825	-0,833
CLUSTER_2_VLC	-0,933	0,673	-0,877	0,714	**0,892**
CLUSTER_3_VLC	-0,528	-0,010	-0,420	0,596	0,555
CLUSTER_4_VLC	-0,934	0,804	-0,913	0,609	0,923
CLUSTER_5_VLC	0,858	-0,428	0,757	-0,635	-0,936

The results were explored to build 'metaclusters' combining similar clusters in the three cities. The criterion was to maximise the average coefficient of correlation for Madrid-Seville, Madrid-Valencia and Seville-Valencia. Table 8 shows the metaclusters proposed with the average correlation coefficient achieved. Metaclusters D and E are less robust than the first three.

Table 3.8: Age metaclusters and coefficients of correlation between their clusters

	Cluster MAD	Cluster SVQ	Cluster VLC	SVQ-VLC	MAD-SVQ	MAD-VLC	Avg
Cluster A	2	5	2	0,892	0,906	0,904	0,901
Cluster B	5	1	1	0,981	0,939	0,708	0,876
Cluster C	1	2	4	0,804	0,836	0,956	0,865
Cluster D	4	4	3	0,596	0,614	0,199	0,470
Cluster E	3	3	5	0,757	0,288	0,082	0,375
				4,030	3,582	2,850	3,487

Figure 3.12 shows the clusters grouped in metaclusters A to C. The following aspects can be highlighted:

- The age distribution of metaclusters A and C are fairly similar, although metacluster C is less conclusive as at least 10% of the cases fall behind each group.
- Metacluster A is likely to correspond to workers with rather high commuting distances, given that it is concentrated in the active population age ranges. This should be confirmed looking at the descriptive statistics of the related clusters (distance, labour status). This should provide insights about the differences with the clusters grouped under metacluster C.
- Metacluster B is formed by those clusters grouping the youngest and eldest age groups, probably with low average trip distances and low number of trips. It seems to incorporate inactive and unemployed adults. This should be confirmed looking at the descriptive statistics of the related clusters (labour status).

The conclusions obtained imply the need for the following further tests on the metaclusters:

- Look at the descriptive statistics of metaclusters A and C, to interpret why metacluster C includes at least 10% of the cases from all age groups with a similar structure to metacluster A. Labour status may be particularly enlightening.
- Look at the descriptive statistics of metacluster B, to interpret if this can be representative of the 'inactive+unemployed' population. Once again labour status seems to be clarifying.
- Trip distances, the number of trips and other mobility indicators differ substantially with the urban density around home location, as larger trips are needed to satisfy certain needs in low density developments. This may add explanatory power to the experiment as it seems that the needs of certain age groups (e.g., school) are more likely to be found in disperse residential areas than others (e.g., work).

Figure 3.12: Distribution of age groups across clusters in metaclusters A to C

Descriptive statistics of the metaclusters

As suggested above, the descriptive statistics of metaclusters A, B and C were studied to see if their age and gender distributions could be explained by the following factors:

- The distribution of the numerical variables presents in each cluster.
- The distribution of labour status present in each cluster.

After applying several tests, the highest differences in the numerical variables were seen in the average distance, although the radius of gyration had a similar distribution. The distribution of the average distance for the three metaclusters is shown in Figure 3.13, Figure 3.14 and Figure 3.15. The following conclusions can be extracted:

- Individuals from metacluster A travel distances 2 standard deviations higher than the average. It is important to note that the distributions plotted represent the distance variable after applying the standardisation. Therefore, in these types of clusters the algorithm is being able to group people who travel long distances.
- Individuals from metacluster C also travel higher distances than the average, especially for Valencia and Seville, but not as much as metacluster A.
- Individuals from metacluster B behave in a whole opposite way to the other two metaclusters. People from this metacluster travel less daily kilometres than the average respondent, which gives us a hint that this group is formed in its majority by the 'inactive + unemployed' population.

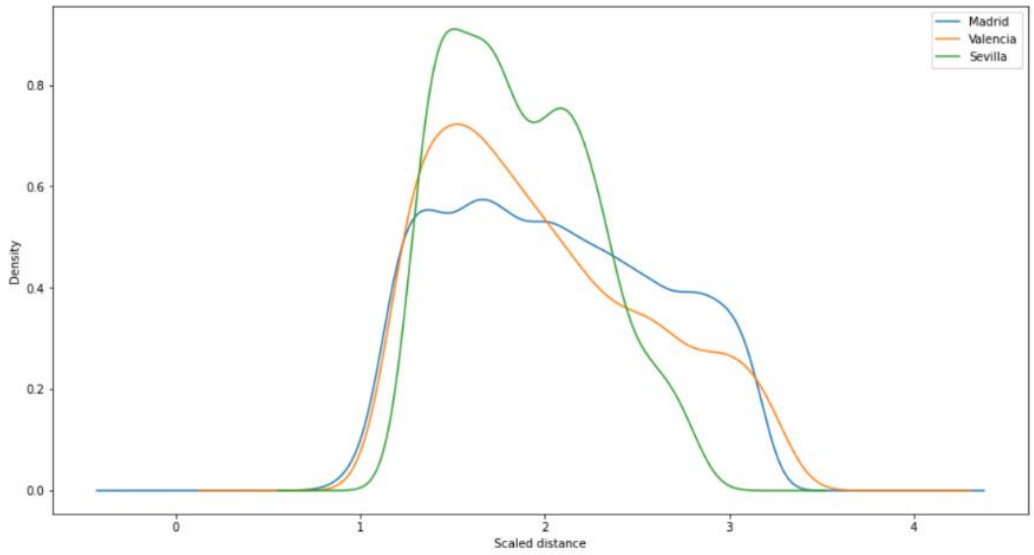


Figure 3.13: Distance distribution for metacluster A in the three surveys.

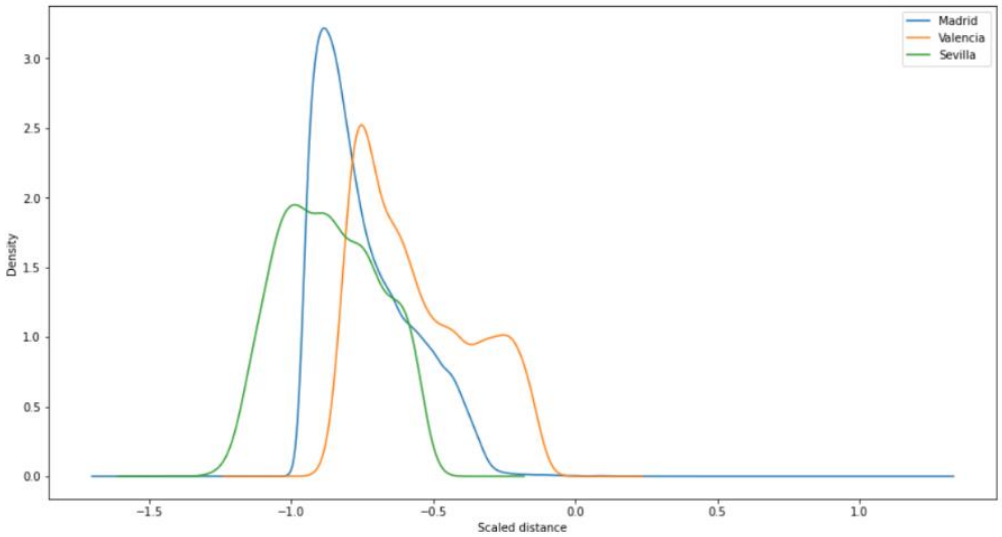


Figure 3.14: Distance distribution for metacluster B in the three surveys

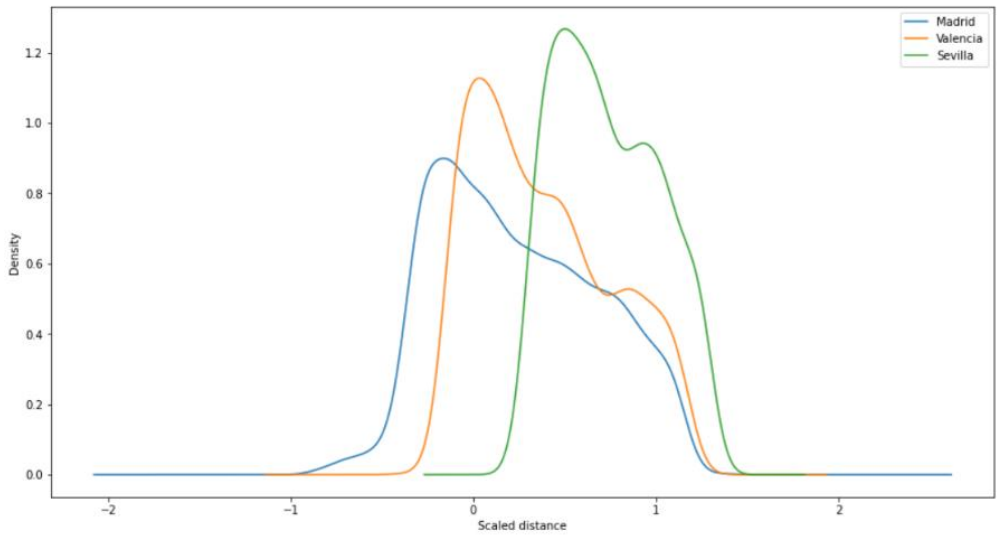


Figure 3.15: Distance distribution for metacluster C in the three surveys

The labour status distribution of each metacluster was also computed and it was compared to the labour status distribution of the whole sample. The comparison made for Madrid's survey is shown in Figure 3.16, Figure 3.17 and Figure 3.18. The following conclusions can be extracted from the plots:

- Regarding metacluster A, in the three surveys the percentage of people whose labour status is 'Work' is higher than the percentage of people that work in the whole survey. This suggests that this group may correspond to workers with high commuting distances.
- Metacluster C behaves similarly to metacluster A in terms of labour status, although the difference with respect to each of the controls is lower in this case.
- Labour status distribution for metacluster B shows that the percentage of people inactive or unemployed is higher than the percentage for the whole corresponding survey. This supports our hypothesis that this metacluster is mainly formed by 'inactive+unemployed' population.

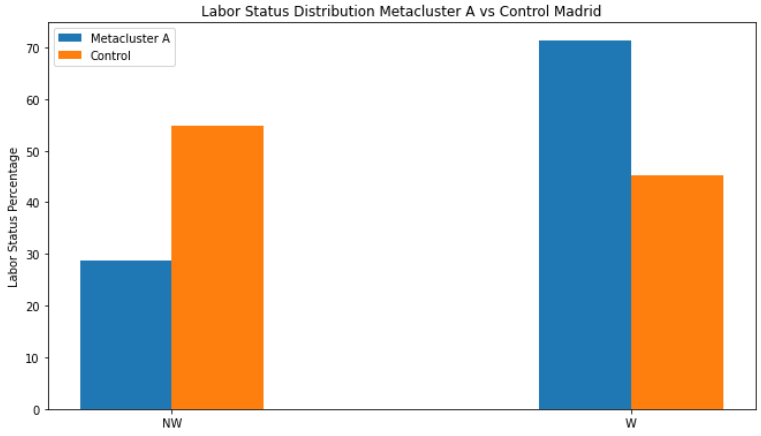


Figure 3.16: Labour status distribution for Madrid's survey in metacluster A

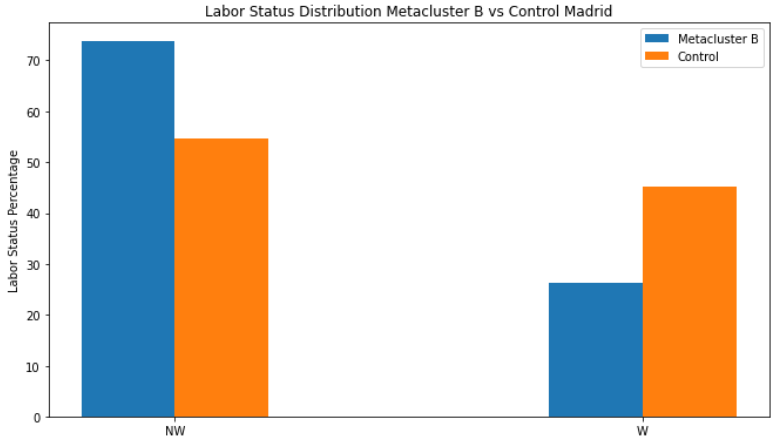


Figure 3.17: Labour status distribution for Madrid's survey in metacluster B

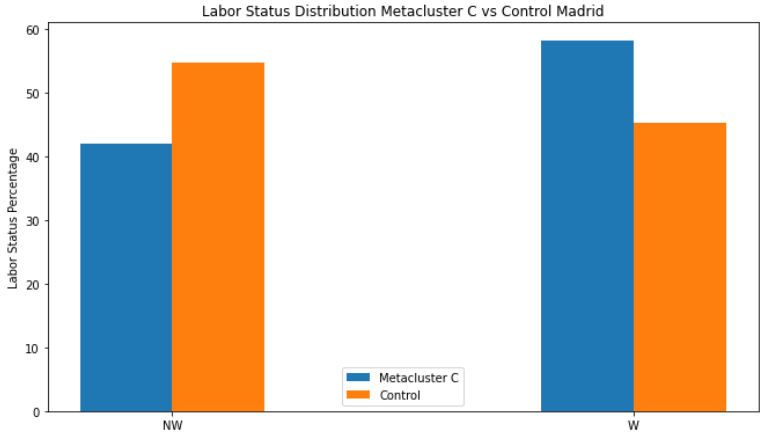


Figure 3.18: Labour status distribution for Madrid's survey in metacluster C

Influence of urban density on travel distances of different age and gender groups

Urban density around home location is known to have a heavy influence on mobility patterns, as larger trips are needed to satisfy certain needs in low density developments. Density varies a lot across metropolitan areas (Figure 3.19). If the effect of urban density on trip distances were equal across all age and gender groups, this effect would not be relevant in this context. However, it may be the case that the needs of certain age-gender groups are more likely to be satisfied in disperse residential areas than the needs of other age-gender groups. A plausible example is school vs. work: home-work distance seems to increase more sharply in low density developments in comparison to home-school distance. The objective of this test was to analyse to what extent the relation between trip distance and sociodemographic group is mediated by urban density. This could help improve the distance-based clustering methods and understand their limitations.

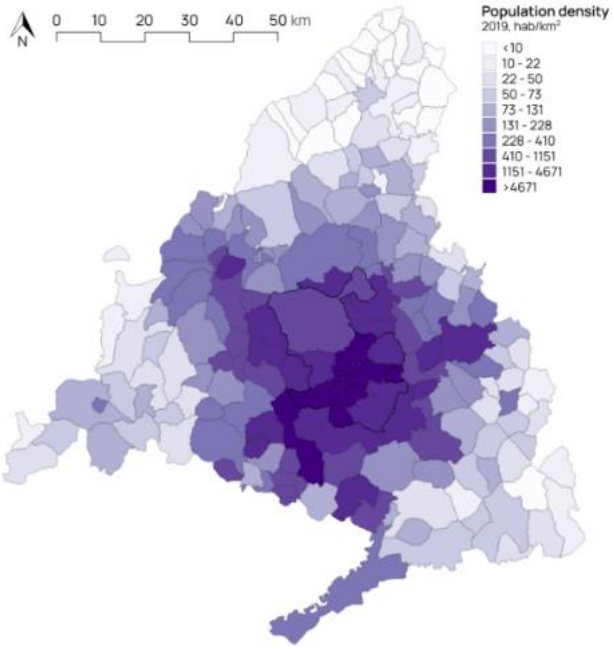


Figure 3.19: Population density in Madrid region

First, trip average distances and urban density were analysed at municipality level for the Madrid region and district level for the Madrid city. Only spatial units with more than 30 sample trips were included and those trips above 200km were removed from the sample, as the analysis focused on metropolitan trips. Urban density accounted for a significant proportion of the average trip distance variability across the region (R^2=0.61), as can be seen in Figure 3.20.

Second, the relationship was analysed across different age groups to evaluate if it varied a lot depending on the age (Table 3.9). The results showed that the youngest and eldest groups present a weaker relation. There are fewer samples available from these groups, as can be seen in the percentage of spatial units that fall above the sample size threshold. However, it seems that this is not fully responsible for the weakness or robustness of the relation, since age groups with similar sample availability (e.g., 0-10 range and 20-30 range) have very different R^2. These results suggest that the hypothesis about the differential relation of urban density and trip distance depending on the age was correct.

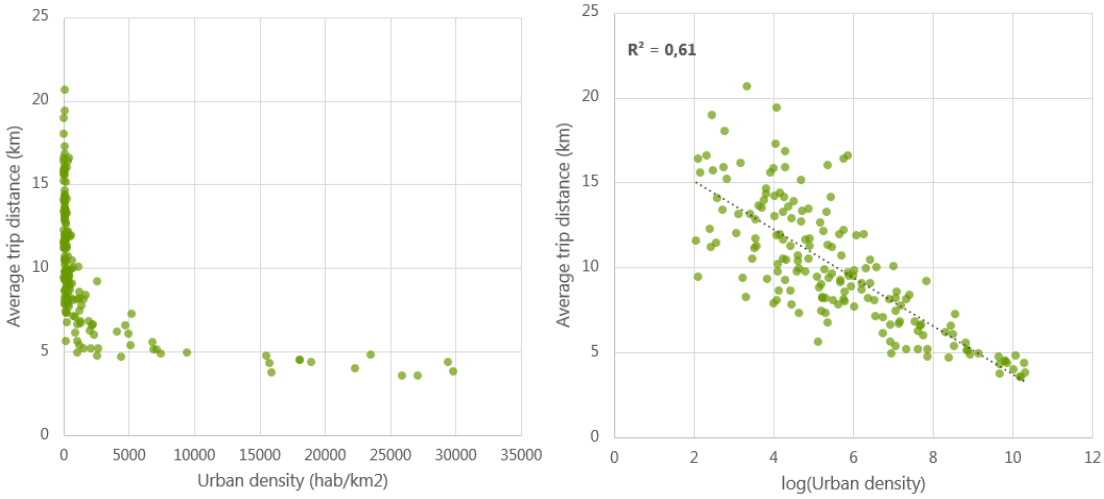


Figure 3.20: Population density in Madrid region (municipalities and districts in Madrid city)

Gender itself does not have a great influence on the relation between urban density and average trip distance. The coefficient of determination is 0.55 for men and 0.53 for women. However, when age-gender categories are defined, different trends can be observed. Following the results of the age analysis, categories related to the working age population were established. The results can be seen in Table 3.10. While male travellers' trip distances are highly related to urban density among the working age population, this is not the case for female travellers. It must be noted that the available units for establishing the relationship is lower in this age group compared to the others.

Following these results, it seemed interesting to add urban density as a clustering variable, taking into account how this mediates in the relation between trip distances and age-gender traveller profiling.

Table 3.9: Relationship between urban density and average trip distance in Madrid region across different age groups. The percentage of spatial units (municipalities/districts) included given the sample threshold (n=30) is shown.

Age range	R^2	% units above threshold
0-10	0,15	31,6%
10-20	0,34	52,6%
20-30	0,51	36,7%
30-40	0,42	39,3%
40-50	0,37	61,2%
50-60	0,47	63,8%
60-70	0,47	41,8%
70-80	0,22	27,0%
80-90	0,09	19,4%

Table 3.10: Relationship between urban density and average trip distance in Madrid region across different age-gender groups. The percentage of spatial units (municipalities/districts) included given the sample threshold (n=30) is shown.

Male			Female		
Age range	R^2	% units above threshold	Age range	R^2	% units above threshold
0-17	0,26	50,6%	0-17	0,23	40,0%
18-44	0,37	52,4%	18-44	0,39	61,8%
45-64	0,51	60,6%	45-64	0,36	74,7%
>65	0,33	32,9%	>65	0,47	31,8%

Addition of urban density to the clustering algorithm

Following the results shown in the previous experiment, the variable representing the urban density of the home municipality (or district in the case of Madrid city) was added to the clustering method for the survey of Madrid. The other variables used in the clustering were the number of trips, the average distance and the radius of gyration.

The silhouette scores obtained from this experiment compared to the ones obtained in the first clustering (without adding density) are shown in Table 3.11. It can be seen that there is a decrease in the value of the scores for every value of the number of clusters *k*. More concretely, the best score obtained is 0.39, 7 points below the best score without adding the density as a clustering variable.

Table 3.11: Silhouette score for each survey after removing "peak" and "off-peak" variables.

	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10
No Density	0.44	0.46	0.43	0.46	0.46	0.43	0.44	0.45	0.44
Density	0.34	0.36	0.38	0.39	0.36	0.38	0.39	0.39	0.35

These results suggest that the 'urban density' variable does not add any information that the k-means algorithm can use to separate the samples into clusters, so this variable was discarded for the next step of the process: the application of k-means clustering to the mobile phone data users.

3.2.3. Clustering on mobile phone data users

After developing an approach to cluster users in terms of their mobility patterns on survey data, the same approach was applied to the agents obtained from the mobile phone data sample. However, as mentioned in Section 3.1, the majority of the mobile phone data users do not provide reliable information of their age and gender. Therefore, in order to be able to analyse the clustering results in terms of the age and gender distribution of the clusters obtained, the sample of the users with reliable values for both characteristics must be obtained. This was accomplished in three steps:

- First, the client portfolio of the mobile phone data records was processed to obtain a mapping relating the contract ID to the age and gender of each user under that contract.
- After analysing this relationship, two issues that confirmed the lack of reliable information for both variables were identified:
  - There are users without age and gender information under their contracts.
  - When there is more than one member under the same contract, the ages and genders of all the members are duplicated.
- Therefore, the reliable sample computed was only formed by the users that have age and gender values and are the sole members of their contract. The final size of the reliable sample was 5.3 million users out of the total 17 million users available for the study period selected, which was the first four weeks of October 2019.

The three variables computed were the same ones obtained from the surveys: number of trips, radius of gyration and distance. However, as opposed to the survey, in this case the variables were calculated for a higher number of days (the 28 of October 2019 mentioned), to take advantage of the longitudinal power of mobile phone data. After that, the average values were computed to obtain the final input variables.

The clustering procedure followed was K-Means clustering, and the possible values for the number of clusters *k* tested ranged from 2 to 7. The *silhouette scores* obtained can be seen in Table 3.12

Table 3.12: Silhouette score values obtained after applying clustering on Orange data

	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7
Orange Data	0.48	0.38	0.39	0.34	0.34	0.35

From the previous table the following conclusions can be obtained:

- In general, the silhouette scores are lower than in the surveys, where the worst results for *k*=5 were around 0.45.
- The best segmentation in terms of the silhouette score is *k*=2, which means that the clustering has not been able to separate enough the sample for higher values of *k*.
- As *k*=2 and *k*=3 are quite low values for the number of clusters, the final value of *k* selected was *k*=4, to see if there were any patterns that could be identified for different groups of age and gender, as in the surveys.

As it can be seen in Figure 3.21-Figure 3.24, the cluster in which there is a biggest separation between age groups is cluster 4, where the majority of the population from elderly groups is represented. Despite this fact, the rest of the group ages have also a high percentage of members in this cluster (>40%). Cluster number 3 looks like the opposite to cluster 4, composed of people from all groups, but with a higher representation of young people. Overall, there are similarities between the tendencies observed in these clusters and the clusters obtained from the surveys. However, the silhouette score is not high enough for k > 2, and there are clusters (e.g., cluster 1) where there are almost no differences between most of the groups.

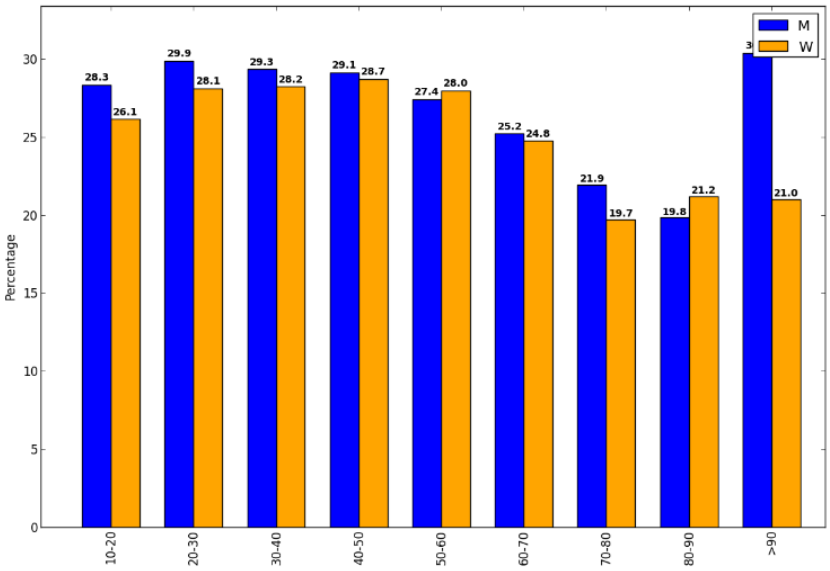


Figure 3.21: Age and gender distribution for Cluster 1

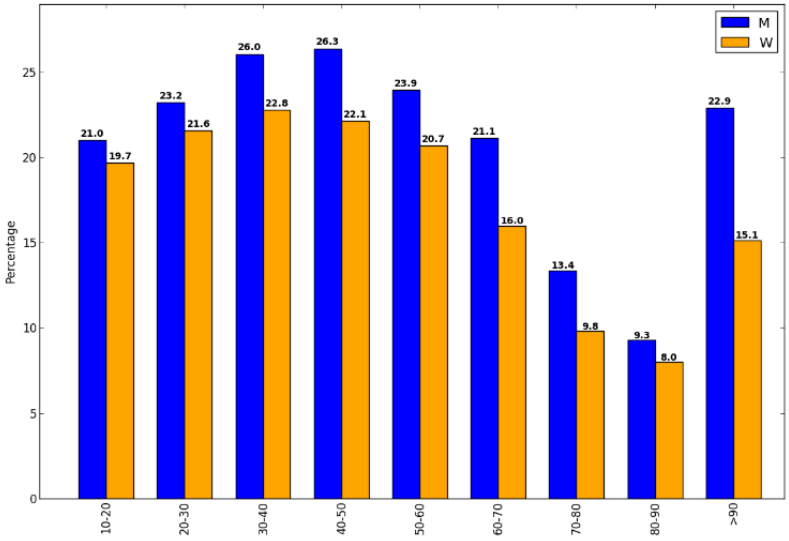


Figure 3.22: Age and gender distribution for Cluster 2

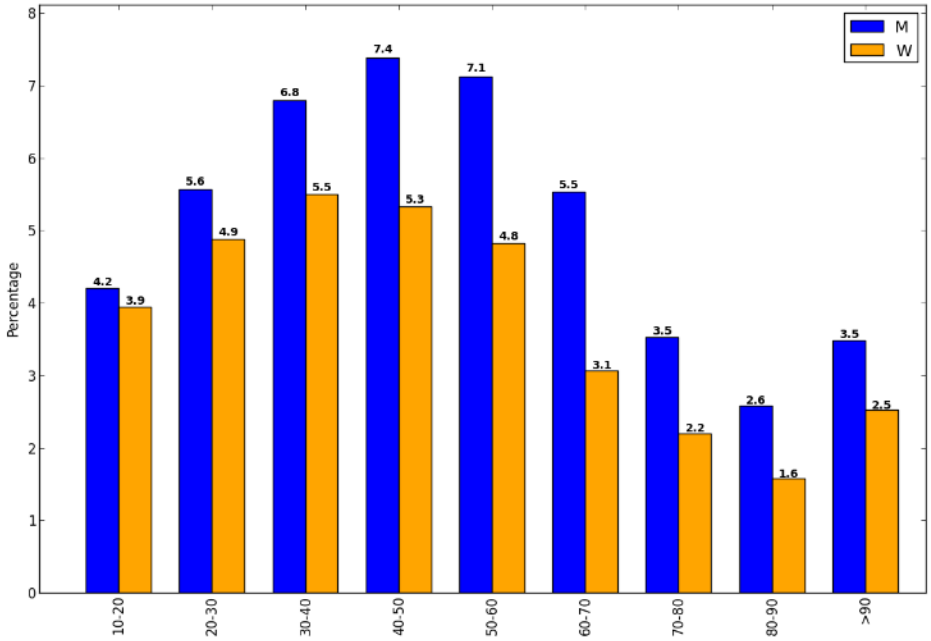


Figure 3.23: Age and gender distribution for Cluster 3

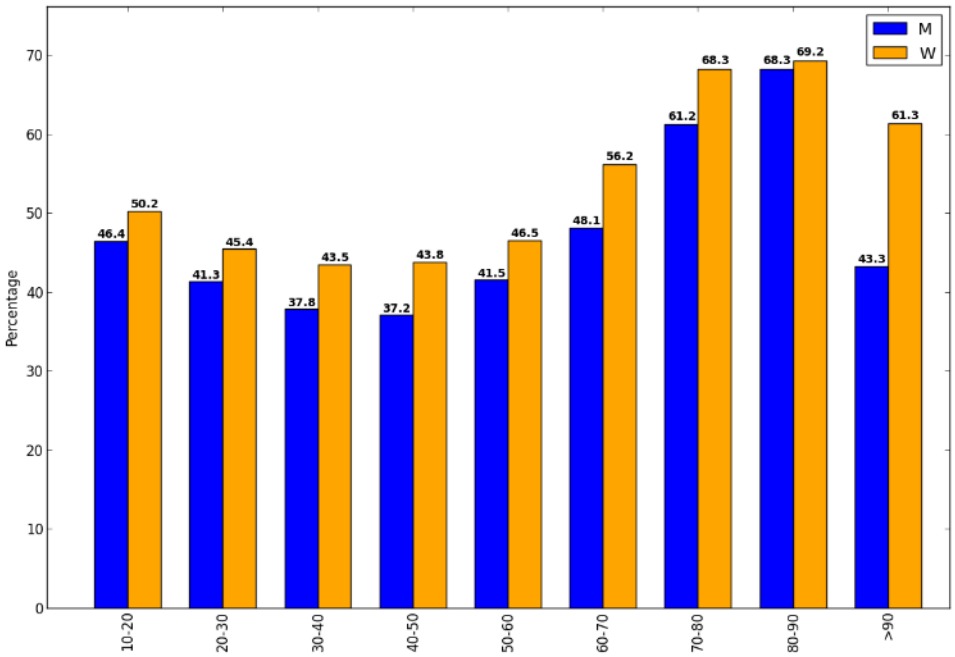


Figure 3.24: Age and gender distribution for Cluster 4

After observing the results for both the surveys and the mobile phone data, it was clear that the silhouette score returned generally low values and the clusters did not seem to separate well different groups of age and gender for some clusters, so it appears that assigning age and gender probabilistically from the distributions of these clusters would not be the most appropriate method.

3.2.4. Supervised learning approach: estimation of age and gender

Once it was clear that the clustering approach did not yield the results expected, the focus was changed into a supervised learning approach. The objective of this new method consisted in developing a machine learning (ML) algorithm able to estimate the age and gender of the individuals that do not have that information or that have incorrect values assigned to them. The general methodology of the solution consists of the following steps:

1. First, two large samples are calculated: the reliable sample (used to train and test the predictive model) and the evaluation sample (all users that are going to be assigned a new age and a new gender). The reliable sample is the same one that was introduced in Section 3.2.3, extracted with the purpose of obtaining only those users with reliable values for age and gender. The evaluation sample is further split into the "evaluation sample for age" and the "evaluation sample for gender", as some users might belong to only one of the two groups.
2. The mobility patterns that are going to be used as input for the predictive model have to be calculated for the two sets. In addition to the variables used in the clustering tests, in this step we introduce a couple of work-related indicators, to better capture active population compared to the rest of the users. In addition, the number of trips and distance variables are split into four subgroups: for week days, weekends and short and long distance. As in the clustering model, the users of the mobile phone data sample with their mobility patterns have been extracted for 28 days of October 2019.
3. Once the input variables were calculated for the users in the reliable sample, the focus was put on developing the ML models able to estimate their age and gender. In terms of the type of ML problem, the gender assignment problem (men or women) is a classification problem. On the other hand, age can be treated both as a numerical or categorical variable (i.e., divided into several groups). The second approach was thought to be more suitable for our purposes, as, for example, there is not a big difference in the mobility between people aged 40 and people aged 45. The final classes selected were: 0-18, 19-44, 45-64 and >65. However, before training the models, it was found that the reliable sample did not include any users younger than 18 (all the users that are single members of their contract are older than 18). To overcome this issue a nested approach for predicting age was followed. First, a ML model was trained using the mobility patterns from Madrid's household survey to classify users in two groups: under-age and adults. Next, a ML model was trained using the reliable sample from mobile phone data to further classify adults into smaller age groups (19-44, 45-64 and >65). To sum up, the selected approach was the following:
    3.1 Build the age model:
        - Build a ML model trained on survey data to predict whether a user is above or below 18 years old.
        - Build a ML model on the reliable sample users to estimate whether their age is in the group 19-44, 45-64 or >65.
    3.2 Build the gender model: train a ML model able to estimate gender.

4. Once the models were trained, the users in the "evaluation sample for gender" were classified as a man or a woman, and the users in the "evaluation sample for age" were assigned an age group. As the final synthetic population built needs to have a numerical age value, each user was randomly assigned an age in their predicted group.

Computation of the reliable sample

The first step of the methodology is to obtain the sets of users that have reliable values of age and gender and the individuals that do not. The procedure to compute these samples is the following:

- The reliable sample of users is computed as explained in Section 3.2.3, using as input data the Orange client portfolio.
- The users without reliable age and gender information are divided into two groups: the evaluation sample for age and the evaluation sample for gender, as there might be users with correct values of only one of the characteristics. The size of these samples is 11.7 and 11.6 million users, respectively.
- The output of this process consists of three sets: the reliable users, with their age and gender, the users with incorrect age, and the users with incorrect gender.

Extraction of mobility patterns

Once the three sets of users were calculated, the input variables for the ML models must be computed for the users in all the sets. There are two types of mobility variables that have been used as input variables to the models: work and non-work indicators. The non-work indicators include the variables already used in the clustering process: average number of trips, average distance per trip and radius of gyration. However, for both the number of trips and the distance, four different values were calculated, depending on whether the trips were on weekdays or the weekend, and depending on whether the trips were short or long-distance:

- Long-distance trips were defined as the trips where the origin province is different from the destination province. Although there might be short trips between provinces, this pattern tries to capture those users that have to move to other regions mainly because of work purposes. After several experiments it was clear that it adds more predictive power to the model than simply using trips above a given distance threshold.
- As seen in the clustering analysis, there are young users that behave as elderly people when it comes to mobility. More concretely, the results obtained in the clustering exercise pointed to work-related reasons as an explanation for the similar mobility patterns of some young and elderly users. Hence, to distinguish these two groups we decided to add mobility patterns for weekends, which are believed to be very dissimilar between them.

Regarding work indicators, the ones selected have been the average number of work trips and the home-work distance. Except fro the home-work distance, the rest of the mobility patterns selected are obtained as an average of all days of study (e.g., the number of trips) or as an average per trip (e.g., the distance).

Table 3.13 shows the eleven variables selected in the creation of the dataset. Once all the variables have been calculated for all users in the three sets, the input data is ready for the next step of the process: the application of the ML model that estimates the age and gender values.

Table 3.13: Mobility patterns selected for the creation of the dataset

Variable	Segmentation	Source
Average number of trips	Short distance in weekdays.	Literature
	Long distance in weekdays.	Mobile phone data
	Short distance in weekends.	Mobile phone data
	Long distance in weekends.	Mobile phone data
Average distance per trip	Short distance in weekdays.	Literature
	Long distance in weekdays.	Mobile phone data
	Short distance in weekends.	Mobile phone data
	Long distance in weekends.	Mobile phone data
Radius of gyration	-	Literature
Average number of work trips	-	Mobile phone data
Home-work distance	-	Mobile phone data

Creation of the ML age model

As mentioned above, one of the problems of the mobile phone data sample is that all reliable users (i.e., all users that are the only members of their contract) are adults, so there is no sample of people below 18 years old. Because of that, another way to assign a value below 18 to an important percentage of the mobile phone data population was sought. The approach followed is based on the use of survey data:

- The idea consists in building a ML 2-class classification model trained using the survey data to predict whether a user is above or below 18 years old, and then evaluate it on the whole "evaluation sample" of the mobile phone data to get the set of users under 18 years old. The survey selected was Madrid's 2018 household survey, which was the one with the greatest granularity and the highest number of samples out of the surveys presented in Section 3.2.1.
- To be able to apply a model calibrated on a survey to the mobile phone data sample, the input variables of the model must be computable from both datasets. Because of that, the number of variables we can calculate is lower than the ones mentioned in the previous section, as longitudinal information like long-distance trips and data over the weekends cannot be extracted from the survey.
- The final variables used to calibrate the model were the average number of short-distance trips in weekdays, the average distance in those kinds of days, the radius of gyration, the number of work trips, and the home-work distance.

- The output value that the model has to estimate, as mentioned above, is a categorical variable with two possible values: 0-18 and >18.
- The procedure followed to calibrate the model consisted of the following steps:
    - First, the dataset is divided (stratified) into a training and a test set. The split used was 80% of the samples for training and 20% for test.
    - Then, a model selection process is applied. During this process, several ML techniques, with different combinations of their parameters, are run on the training set to calibrate a model that is able to estimate whether a survey participant is an adult or not. For each one of these techniques, to obtain the best model, the *k*-fold cross-validation resampling technique is used. This method separates the training data into *k* stratified portions, and trains the model with *k* times using as training sample *k-1* of those groups, and validates the result on the remaining portion. The number of folds or groups used in the cross-validation was set to *k*=5, as it leads to an 80/20 partition of the training data, which seems appropriate having in mind the size of our dataset.
    - After computing the average validation error for every technique and every combination of parameters, the best model is selected as the one with the least average error. The ML algorithms tried were decision trees, random forest, support vector machines and neural networks (multilayer perceptron). The results are shown in Section 3.4.
    - Once the best model with the best parameters is selected, it is trained using the whole training set and it is evaluated on the test set that was left behind in the first step, to obtain a final indicator on how the model behaves with unseen data.
- After the final model is trained, it is applied on the "evaluation sample for age" and it assigns to each user of the sample one of the two possible labels mentioned above: 0-18 or >18. The users who have been assigned the label 0-18 are assigned a random age in that interval and are never used again in the next steps of the process. The rest of the users, on the other hand, are going to be further segmented by applying a new ML model, which is explained next.

Once the problem of being able to distinguish between under-age and adult users is solved, the mobile phone data can be used to predict the age of the adult users. The main steps applied for building the model in charge of doing it were the following

1. Define the output variable: as the age can be treated as a numerical value or grouped into different segments, there are two main possibilities: treat this as a regression or as a classification problem. The second approach was thought to be more adequate for our purposes. Two different divisions in groups were tested:
    1. 19-24, 25-44, 45-64 and >65.
    2. 19-44, 45-64 and >65.
2. Obtain the input variables needed for every user in the reliable sample, to be able to train the model on those users. This step has already been explained above, and its output consists of a dataset that contains the 11 numerical mobility variables which are going to be used as input to the model. As mentioned above, the variables have also been calculated for the users that need to be evaluated on the model.
3. Select the training strategy. Apart from the training-test stratification and the cross-validation procedure explained for the survey model, which were the same in this case, an important aspect to bear in mind is how to approach the fact that the sample is imbalanced in the case of

the age model, as some of the groups are smaller than the rest. In order for the models to be trained properly, it is needed to have a representative and well-balanced sample, and in this case, only 6% of the reliable users in the 19-24 group, 43% are between 25 and 44, 36% belong to the 45-64 group, and 15% are older than 65.

4. Select the ML models to apply in each case. A range of ML models were tested: random forest, gradient boosting, neural networks and KNN. As in the case of the survey model, a wide range of parameters for each model was tested to obtain the best possible model. Once the best model was picked, it was trained using the whole training set and a final indicator of its evaluation on the test set was given.

Regarding the first step, in Section 3.4 the results for the two options tested for the age groups are shown. Analysing those results, a decision was made on which partition to use in the final model. On the other hand, as step 2 has already been explained in previous sections, in the rest of this section the main focus will be placed on steps 3 and 4.

Training approach

One of the most important aspects when building a ML model is to define a training strategy in which to test multiple models and select the one that best fits the output variable. As done in the survey model, the first steps applied when building the model were the division of the mobility dataset into train and test and the application of cross-validation to select the models that best predicted age and gender. The number of folds for the cross-validation selected was again 5, as in the survey model.

Once the partition into training and test and the cross-validation procedure were performed, the first problem encountered when calibrating the model was, as mentioned above, the fact that the number of samples belonging to each of the age groups in the reliable sample was different, especially higher in medium ages. Therefore, it was clear that some approach to balance the input dataset had to be applied. There are two typically known techniques for solving imbalanced problems:

- Downsampling: this technique reduces randomly the size of all the groups but the smallest one, so that the size of the sample for each group is the same. The main drawback of this method is that a lot of information is lost when removing the individuals of the bigger groups.
- Oversampling: this method is the opposite of the previous one. It randomly duplicates elements from the smaller groups to match the size of the biggest one, so that again the size of the sample for each group is the same. This method does not lose any information by removing elements but it has the problem that some of the samples are repeated and do not provide any additional information to the model. One of the most known oversampling techniques is a variation of this random oversampling method, called SMOTE, which basically oversamples the smaller groups by creating new synthetic samples that are not exactly duplicated from the existing ones, but a combination from them.

For our problem, two different approaches were followed: not applying any balancing technique, and using the SMOTE oversampling algorithm. The results of both tests are shown in Section 3.4. The SMOTE technique resulted in a better estimation for the smaller groups, so this approach was selected for the final model.

Another pre-processing procedure that was applied is standardisation, which has been already introduced in the clustering analysis explained in Section 3.2.2. Depending on the ML model, one needs or does not need to standardise the numerical input variables. However, it was decided, as mobility indicators are different depending on the region you live in (e.g., people in Madrid travel in average 7 km per day while people in Valencia travel 4), that the input variables should be standardised by home district of the user, no matter which model was used.

One of the most typical mistakes when applying these pre-processing techniques (standardisation, SMOTE and cross-validation) is the order of application of each one of them. It is important to take into account that if we create the folds of the cross-validation after applying SMOTE, some of the synthetically created samples might be present in the validation set while their "originals" might be in the training set. Therefore, it would be easier for our model to classify the validation samples as we are introducing noise in it. It is true that the synthetic and their original samples are not the same, but they are similar in a way, so it is important not to apply the oversampling method first. Therefore, the right way is to do the cross-validation, loop over each fold, and apply the SMOTE technique only on the training part of that fold, and leave the validation set as it is. The same idea applies to the order of application of the creation of the folds and the standardisation. The standardisation has to be fit in the training part of each fold (i.e., the mean and standard deviation of each variable are computed on the training part), and then it is applied directly to the validation part.

Model selection

Several ML algorithms have been tried to train the age model using the reliable sample: random forest, gradient boosting, multilayer perceptron and k-nearest neighbours. The experiments to test the performance of each of the algorithms used as input a small random subset of the reliable sample (400K out of 5M), as some of the models were very computationally intensive. To compare the performance of these models, a tuning of their hyperparameters was performed and the cross-validation procedure was used to select the best model (and its hyperparameters) as the one with the highest validation score. The score metric used was the F1-Score, which computes the harmonic average between precision and recall. The definition of each of these metrics is the following:

- **Precision**: Number of users correctly classified as being from one class with respect to the total number of users classified as being from that class.
- **Recall**: Number of users correctly classified as being from one class out of the total number of users that are labelled with that class, i.e., the users that are actually the age or gender the class represents.
- **F1-score**: Harmonic average of precision of recall. As there might be models with, for instance, high recall and low precision, it is interesting to give a unique number that takes into account both metrics. Using the harmonic mean avoids hindering underperformance from one of the two metrics (e.g., it prevents the model to score 0.5 when the precision is 1 and recall 0). The F1-Score was the metric used for making the decision of picking the best model with its best parameters.

Model evaluation

Once each of the best models is obtained, it is used to assign the age to the users that need it. As the model returns the group of ages to which each user belongs, a random numerical value was computed for them within their predicted group, having all possible ages in the group the same probability of being selected. The value obtained represents the final age assigned to each user.

Creation of the ML gender model

Regarding gender, the approach followed was equivalent to the one applied for predicting an age value for the adult agents, with the difference that the former is applied to all the users in the "evaluation sample for gender" and the latter is only used to assign an age to the adult agents. Therefore, this model is not nested, and it only needs mobile phone data to be trained. The input variables selected in this case are also the ones introduced in Table 3.13. Regarding the training procedure, the same pre-processing techniques were applied, but without having to solve the imbalance problem, as there were almost the same number of men (51%) and women (49%) in the reliable sample. Regarding the model selection process, the ML algorithms used were random forest, gradient boosting, multilayer perceptron and k-nearest neighbours. As in the age model, once the best model was computed by selecting the one with the highest cross-validation F1-Score, the users in the "evaluation sample for gender" were assigned one of the two possible values: man or woman.

3.3. Validation plan

3.3.1. Overview

The calibrated model has been subsequently validated to ensure that the model produces sensible results and that the resulting model is not overfitted.

3.3.2. Objectives

Mobile phone data sometimes lacks a good age and gender characterisation. By analysing the travel patterns through ML techniques and household surveys, age and gender can be derived. This enhanced characterisation will be used as an input when it comes to the characterisation of the users' travel decisions.

3.3.3. Validation approach

The validation approach is summarised below:

- Divide the used data into training and test data
- Stratify the training data into different subgroups through the application of a cross-validation procedure. The technique used is *k*-fold cross-validation and consists of the following steps:
    - Divide the training set into *k* portions.
    - For each one of the *k* portions:
        - Use *k-1* groups to form the new training set and the remaining portion to create the validation set.
        - Train the model and evaluate it on the validation set to obtain an estimation of the skill of the model on unseen data.
- Select the best model and its best parameters among the ML models run as the one with the highest average validation F1-Score.
- Use the selected model with the best parameters to train a final model on the whole initial training set.
- Validate the resulting model with the test data.

3.3.4. Data and software inputs

To calibrate and validate the prediction of age and gender covered in the passenger characterisation algorithm, the mobile phone data, which is only available for Nommon, has been used as input. To obtain a sample of mobile phone data users with reliable information of age and gender, an analysis of Orange's client portfolio was performed. In this analysis, two issues were identified:

- There are users without age and gender information under their contracts.
- When there is more than one member under the same contract, the ages and genders of all the members are duplicated.

Therefore, to clean the total sample of users and keep only the reliable ones, only the agents having age and gender information and being the sole members of their contract were selected.

Regarding the parameters used to validate, both for the age and gender models, the validation inputs are the following:

- training test split: 80/20
- cross-validation split: 5 folds

3.4. Results

3.4.1. Age model

As mentioned in Section 3.2.4, the age model was divided into two steps: estimating whether a user is under-age or adult through the use of surveys, and estimating the age of the adult users using mobile phone data. To build the survey model, the techniques used to train the survey model were decision trees (DT), random forest (RF), support vector machines (SVM) and neural networks (Multilayer Perceptron, MLP).

The cross-validation and test results for all the models mentioned above can be seen in Table 3.14The model with the highest cross-validation score is the random forest model, followed closely by the decision trees and the neural networks. In terms of the differences between the classes, all models are able to identify almost perfectly the adult participants, while the individuals from 0-18 are harder to predict. However, having in mind the differences in sample size between both groups, the results are satisfactory. Although it seems like a good opportunity to apply a balancing technique, like SMOTE, to correct the balancing problem between the classes, it was seen that it did not improve the results and hence it was not used for the final model.

Regarding the algorithm selected as the best to estimate which users are adult or not, although the random forest yields better results, the decision tree model has been chosen, as it is always better to select a more basic model if the results do not decrease significantly.

Table 3.14: Survey model cross-validation and test results.

		F1-Score DT	F1-Score RF	F1-Score MLP	F1-Score SVM	Number of users
Validation	Mean CV score	0.87	0.88	0.87	0.84	N/D
Test	0-18	0.63	0.64	0.64	0.53	2449
	>18	0.92	0.93	0.93	0.92	10997
	Weighted Average	0.87	0.88	0.54	0.85	13446

Once the model has been calibrated, the last step of the process is to evaluate the output of the mobile phone data users. As mentioned above, the users predicted to have an age below 18 are assigned a random value between 0 and 18. The rest of the users pass on to the next step of the process: the building of the model that assigns an age value to adult users.

Below we explain the different decisions made while training (e.g., the oversampling technique to apply). After that, the comparison of the different models is presented, together with the results of the final model both for age and gender.

Training

As explained in Section 3.2.4, for the age model two possible approaches were tested regarding the balancing problem: not applying any balancing technique (in case the sample is separable enough), and using the SMOTE oversampling algorithm. The results obtained from the application of both approaches are summarised in the following points:

- When using 4 groups as output for the age class, the cross-validation and test results using the random forest algorithm are shown in Table 3.16. It is worth noting that for the pre-processing tests, the model used was random forest, as it did not matter the performance 'per se' of the model, but the selection of the best measure. It can be seen that when the SMOTE procedure is not applied, all the predictions given to the smallest group (19-24) are missed. When rebalancing with SMOTE, on the other hand, the performance is more balanced across all groups. Although in the first case, it is adequate to obtain a F1-Score of 0.66 in the 25-44 group compared to 0.42 with the SMOTE algorithm, it is much worse to have a group in which no users have been estimated properly. In addition, the validation results are better when applying SMOTE.
- When using 3 age groups, the results are more similar between both approaches, as shown in Table 3.16. However, when using SMOTE a better prediction for both of the groups 45-64 and >65 is obtained, smaller in size than the group 25-44, which is more important for our interests than only predicting correctly one of the groups.

Table 3.15: RF model on the test set comparing the performance of the SMOTE technique vs not applying a rebalancing technique when there are 4 age groups.

		F1-Score No Oversampling	F1-Score SMOTE	Number of users
Validation	Mean CV score	0.37	0.40	N/D
Test	19-24	0.00	0.08	4,826
	25-44	0.66	0.53	35,562
	45-64	0.21	0.34	29,077
	>65	0.07	0.36	11,848
	Weighted Average	0.37	0.41	81,313

Table 3.16: RF model on the test set comparing the performance of the SMOTE technique vs not applying a rebalancing technique when there are 3 age groups.

		F1-Score No Oversampling	F1-Score SMOTE	Number of users
Validation	Mean CV score	0.46	0.47	N/D
	19-44	0.64	0.61	40388
	45-64	0.29	0.34	29077
	>65	0.26	0.36	11848
	Weighted Average	0.46	0.48	81313

Although the weighted results for both methods are very similar it is preferable that the smaller classes can be predicted with the highest F1-Score as possible (and, of course, it must not be 0). Therefore, the approach picked was to apply SMOTE in the training of the age model. After seeing the best procedure regarding the balance of the classes, the selection of the optimal number of classes for the "Age" variable can be discussed. As shown in Tables 3.14 and 3.15, the results for the SMOTE technique between both tables are similarly proportional to their number of classes. However, the group containing people between 19 and 24 has quite a poor F1-Score when using 4 age groups. More concretely, its precision and recall values obtained were 0.09 and 0.07, respectively. Because of that, the number of classes selected was three (19-44, 45-64 and >65).

Model selection

The results for the best model selected, i.e., the one with the hyperparameters that got the best mean F1-score through all the validation sets of the cross-validation, for the algorithms random forest (RF), gradient boosting (GB), multilayer perceptron (MLP) and k-nearest neighbours (KNN) can be seen in Table 3.17. The first row of the table represents the mean validation score for each model across the five folds of the cross-validation. The following rows represent the F1-Score for each class in the test set, together with the number of agents present in each class.

The following conclusions can be extracted:

- The age model with the best mean CV score is the random forest (RF). Although it has similar test scores as gradient boosting (GB), the selected model has been RF as it has proved during the cross-validation that it could estimate the five validation samples better than the GB method. The rest of the models are significantly worse.
- With respect to the F1-Score values obtained for the age model, although the results are not particularly high, with a final weighted score of 0.47, it is important to have in mind that the baseline model in this case has a F1-Score value of 0.33. Observing the results, it seems that using only mobility patterns does not let us capture all the existing differences between all age groups. Future improvements that will be explored include the addition of SMS and calls information per day of each user as well as navigation data from online searches

The final ML model computed was thus a random forest. The final parameters of the model were 500 trees, a maximum depth of the trees of 50 and a minimum number of samples to split a node of 9.

Table 3.17: Validation and test results for the age models with different ML algorithms.

		F1-Score RF	F1-Score GB	F1-Score MLP	F1-Score KNN	Number of agents
Validation	Mean CV score	0.47	0.45	0.44	0.41	N/D
Test	19-44	0.61	0.60	0.59	0.54	40388
	45-64	0.34	0.32	0.25	0.33	29077
	>65	0.36	0.39	0.38	0.29	11848
	Weighted Average	0.48	0.47	0.44	0.43	81313

3.4.2. Gender model

The results obtained for the model in charge of estimating gender are shown in Table 3.18.

- All the models except from KNN are quite similar, having the three of them the same cross-validation scores. The model with the highest weighted test score is gradient boosting. However, as the CV scores are the same as for the RF method and the difference in the test set is only by 1 point, the random forest method was selected so as to keep both the age and gender models as much unified as possible.
- In terms of the scores obtained, it seems that gender is more difficult to capture than age using only mobility patterns. One improvement that might be particularly enlightening for this model is navigation data (e.g., number and type of searches made by each user).

Table 3.18: Validation and test results for the gender models with different ML algorithms

		F1-Score RF	F1-Score GB	F1-Score MLP	F1-Score KNN	Number of users
Validation	Mean CV score	0.54	0.54	0.54	0.52	N/D
Test	Male	0.55	0.51	0.50	0.52	41,740
	Female	0.53	0.59	0.58	0.52	39,573
	Weighted Average	0.54	0.55	0.54	0.52	81,313

3.4.3. Validation

To check that the whole supervised learning approach is worth using at the TRANSIT project in future developments and calibrations, it is necessary to compare it to the current profile assignment method developed by Nommon in previous projects. The hypothesis followed by this procedure is based on the relationship between the home-work distance and the age and gender of the users.

After launching this method, its results were compared to the ones obtained for the new random forest model. The results on the test set for both age and gender can be seen in Table 3.19 and Table 3.20, respectively. Looking at the results, it can be seen clearly that for the age value the average F1-Score is lower (11 points) for the current method than for the new predictive model. The only age group that is equally well predicted by both models is 45-64. Another aspect to take into account is that, while the current solution is not able to assign a reliable age below 18 years old (because the data used for the clustering is formed by users whose age might be duplicated), the model developed for this project includes a decision tree trained with survey data that is able to assign under-age people with a F1-Score of 0.63. Regarding gender, despite the fact that the predictive model is not able to capture the complexity of the difference between both groups, the new model also outperforms the previous model.

Therefore, we can conclude that the predictive model implemented gives better results in the prediction of age and gender than the current solution.

Table 3.19: Test age results comparison between the current profile assignment method and the predictive model implemented

		F1-Score Current Component	F1-Score RF Predictive Model	Number of users
Test	19-44	0.43	0.61	40,388
	45-64	0.34	0.34	29,077
	>65	0.27	0.36	11,848
	Weighted Average	0.37	0.48	81,313

Table 3.20: Test gender results comparison between the current profile assignment method and the predictive model implemented

		F1-Score Current Solution	F1-Score RF Predictive Model	Number of users
Test	Male	0.51	0.55	41,740
	Female	0.50	0.53	39,573
	Weighted Average	0.51	0.54	81,313

3.5. Case study: profiling of Madrid-Barajas passengers

After developing a ML model able to estimate the age and gender of the mobile phone users, it is interesting to see how well it is able to estimate the age and gender for the users of the sample that are air transport passengers. To this end, the passengers of any flight arriving or departing from Madrid Barajas have been extracted from the mobile phone data sample, and the ML model developed has been used to assign them an age and a gender value. The data for this study was from November 2018. The following steps were applied to carry out the case study:

- A set of reliable passengers was obtained from the whole passenger sample. These reliable passengers, equally to the reliable sample used to train the model, consisted of passengers having age and gender information and being the sole members of their mobile phone contracts. Out of the total number of 105K passengers, the reliable sample was formed by 43K users. The age distribution of the reliable passenger sample was even more imbalanced than then one of the training reliable sample: 19-44 (55%), 45-64 (39%) and >65 (6%). Regarding gender, no balancing problem was found.
- After obtaining the reliable passengers, the eleven input variables of the mobile phone data model were computed. Before applying the predictive models, the same per-district standardisation process applied to the training data set was applied to the testing sample. This standardised set represents the input passenger set that is going to be the input to the model.
- Once all eleven variables are standardised across the passenger reliable sample, the model is applied on them to make its estimation. The results of the evaluation are shown in Table 3.21 and Table 3.22, for age and gender, respectively.

Table 3.21: Precision, recall and F1-Score values for age evaluation on the passenger sample

	Precision	Recall	F1-Score	Number of users
19-44	0.57	0.65	0.61	24,023
45-64	0.40	0.29	0.33	16,791
>65	0.10	0.18	0.12	2,471
Weighted Average	0.48	0.48	0.47	43,285

Table 3.22: Precision, recall and F1-Score values for gender evaluation on the passenger sample

	Precision	Recall	F1-Score	Number of users
Male	0.57	0.55	0.56	23,807
Female	0.48	0.50	0.49	19,478
Weighted Average	0.53	0.53	0.53	43,285

The following conclusions can be extracted:

- Regarding gender, the results are quite similar to those obtained when validating the model, although male passengers are predicted slightly better than females.
- The overall results for age are also similar to the test evaluation performed when calibrating the model, with a weighted F1-Score value equal to 0.47.
- One of the groups is estimated significantly worse than the others. This segment is the one formed by people over 65 years old. This can be due to two possible reasons:
    - The model is not able to capture the temporal change between the mobility patterns used for the calibration (with data from 2019) from the passengers' patterns (obtained from 2018).
    - The available passengers that are older than 65 are much more active than the majority of elderly people captured in the reliable sample when calibrating the model.

To check that the second option is the reason for the poor performance of the model on passengers older than 65 years and therefore that these passengers are a subset of the users that were not estimated properly in the calibration, the random forest model was evaluated on all the reliable users from 2018. If the results are similar to the estimation obtained while testing on the reliable sample of 2019 while doing the calibration, it is possible to ensure that the model is not underperforming on data from 2018. The results of this evaluation are shown in Figure 3.22.

Table 3.23: Evaluation results on the whole reliable sample from 2018 (including passengers)

	Precision	Recall	F1-Score	Number of users
19-44	0.59	0.53	0.56	923,958
45-64	0.37	0.35	0.36	642,541
>65	0.28	0.44	0.34	244,559
Weighted Average	0.47	0.45	0.46	1,811,058

Analysing the results from Table 3.23, a similar tendency in F1-Score can be noticed in comparison to the test results showed in Table 3.17 while calibrating the model with data from 2019. Regarding elderly users, the score obtained has been only two points lower when evaluating on data from 2018, which implies that the model is not underperforming because of the change of dates between samples, but because the fact the captured passengers over 65 years old behave different than the usual agents that belong to that age group. To confirm this hypothesis, the distribution of the scaled variables for both the reliable sample from 2018 and the reliable passengers from 2018 was computed. It is important to note that the standardisation applied to these samples is the same to the one conducted while training the model. The results of the comparison of the average scaled value for each variable are shown in Table 3.24. All variables are higher for the passenger sample, especially those referring to distance. The most extreme cases are those regarding the average distance in long-distance trips. In weekdays the number of standard deviations above the mean is 1,557, while in weekends it is 570. Having in mind that the variables showed below are scaled, we can conclude that the differences are significant enough to make the model fail in these situations.

Table 3.24: Scaled variables comparison: 2018 reliable sample vs passengers from that sample.

Variable	Segmentation	2018 Reliable sample value	Passenger sample value
Average number of trips	Short distance in weekdays	-0.5	0.01
	Long distance in weekdays	-0.1	0.34
	Short distance in weekends	-0.36	-0.16
	Long distance in weekends	-0.01	0.1
Average distance per trip	Short distance in weekdays	-0.1	45.29
	Long distance in weekdays	-0.09	1557.96
	Short distance in weekends	-0.28	96.88
	Long distance in weekends	-0.09	570.74
Radius of gyration	-	-0.08	9.36
Average number of work trips	-	-0.27	0.18
Home-work distance	-	-0.23	4.06

Therefore, although the random forest model is able to adequately estimate the age and gender of average users of the mobile phone data sample, it underperforms when it has to predict the age of elderly passengers. To solve this, a model dedicated exclusively to the prediction of passengers' age and gender should be developed, selecting as input data the mobility patterns that are most significant when comparing only passengers, instead of all the mobile phone data users.

3.6. Conclusion

In this chapter we have discussed two ML approaches, unsupervised and supervised, for estimation of passengers' age and gender.

The unsupervised ML approach has shown not to be optimal for the objective pursued. However, this approach gave some interesting insights about age related mobility patterns, such as the fact that when looking only to week days, elderly users are mixed up with un-employed people from younger groups, due to their similarity in their mobility patterns.

The supervised approach shows good results when used for age prediction taking as input variables the average number of trips, the average distance per trip and the radius of gyration, being the first two further segmented into weekdays and weekends, and into short and long distance. The ML algorithms with the best results are the decision trees and the random forest.

For the groups covering people below 45 years old, the score obtained is around 0.60, almost duplicating the baseline model, while the F1-Score for the rest of the groups hardly outperforms the baseline model, which is 0.33. However, for people over 65 there is an improvement of 9 points over the current profile assignment solution. On the one hand, the results demonstrate that gender is extremely hard to predict using only the mobility patterns selected as input data.

The results obtained are similar to the ones seen in the few alike methods found in the literature, such as [10], where the F1-Score reached is 0.24 with 6 age groups, and [11] with a final score value of 0.37 over 4 classes. However, there is still much room for improvement in terms of accuracy for the model developed, as other characteristics of the user can be added as predictive variables, such as the ones already mentioned in Section 3.3: average searches of certain online categories (e.g., sports or social networks), data on the number of calls and SMS made and received by each user, etc.

Regarding the case study applied only on passengers from Madrid Barajas airport, it was seen that the model was able to make similar estimations for the passengers below 65 years old, compared to the test results obtained when calibrating the model. However, elderly passengers were poorly predicted as they tend to move differently than the average elderly user, which ends up in much higher values for the scaled variables computed while training the model. For future improvements, passenger users should be separated from the rest, and their sociodemographic characteristics should be estimated by training a model dedicated only to them.

4. Modal choice in airport access

4.1. Problem statement

When using mobile network data to analyse travel demand, segmentation by transport mode usually relies on map matching techniques. However, when it comes to the identification of mode in short trips in urban areas, the higher density of transport modes often makes map matching techniques unreliable. Consequently, another approach has to be considered.

Traditionally, transport demand has been derived using the 4-stage model ([12]), which consists in estimating the demand generated and attracted in each study zone, estimating the distribution of these trips, estimating the mode used for these trips, and finally assigning each mode specific trips to the transport network. To estimate mode choice, logit models are commonly used. In this section we describe the calibration of a mode choice logit model using household surveys, transport supply data and mobile phone data, and how it has been used to estimate modal choices in the access to Madrid-Barajas airport.

4.2. Methodology

The methodology section has been divided into three subsections. In subsection 4.2.1, the methodology followed to obtain the raw transport supply characteristics is explained. In subsection 4.2.2, the methodology followed to refine this transport supply characterisation by adding cost information well as parking availability information is explained. Finally, subsection 4.2.3 illustrates the steps followed in the logit model calibration and validation.

4.2.1. Characterisation of trips

To be able to calibrate a logit model, which will assign a mode to the trips observed from the mobile phone data, the following inputs should be considered:

- Travellers' characteristics which are derived from mobile phone data and enhanced with the ML approach described in Section 3.
- Trip characteristics, which are derived from mobile phone data
- Transport supply characteristics, which are obtained from sources such as Open-Source Routing Machine (OSRM) and OPT (Open Trip Planner).

This subsection describes the transport supply characterisation by explaining how the travel times, travel distances and other transport supply characteristics are obtained for all the considered modes.

As a starting point, only walking, public transport and car are to be considered for the logit model calibration. Transport supply information is to be obtained from OSRM for walking and car and OTP for public transport. The reason for using OSRM and OTP instead of Google Maps is that unlimited free queries can be launched in OSRM and OTP, but not in Google Maps.

Walking travel times (OSRM vs. Google Maps)

To check the quality of the OSRM walking routes, OSRM and Google Maps queries results are compared. OSRM allows the modification of the walking speed, hence the goal of the experiment is to understand the differences between OSRM and Google Maps walking travel times and obtain the OSRM optimal walking speed.

Inputs

The code requesting the route queries includes the following parameters:

- Origin
- Destination
- Walking speeds

Methodology

Almost all walking trips observed in the Madrid household survey (99.5%) have less than 5 km distance (see Figure 4.1). The average trip distance is 740 meters. Walking modal share peaks at very short distances (around 200m), and drops below 50% at around 1,000m and below 20% at around 2,000m (see Figure 4.2).

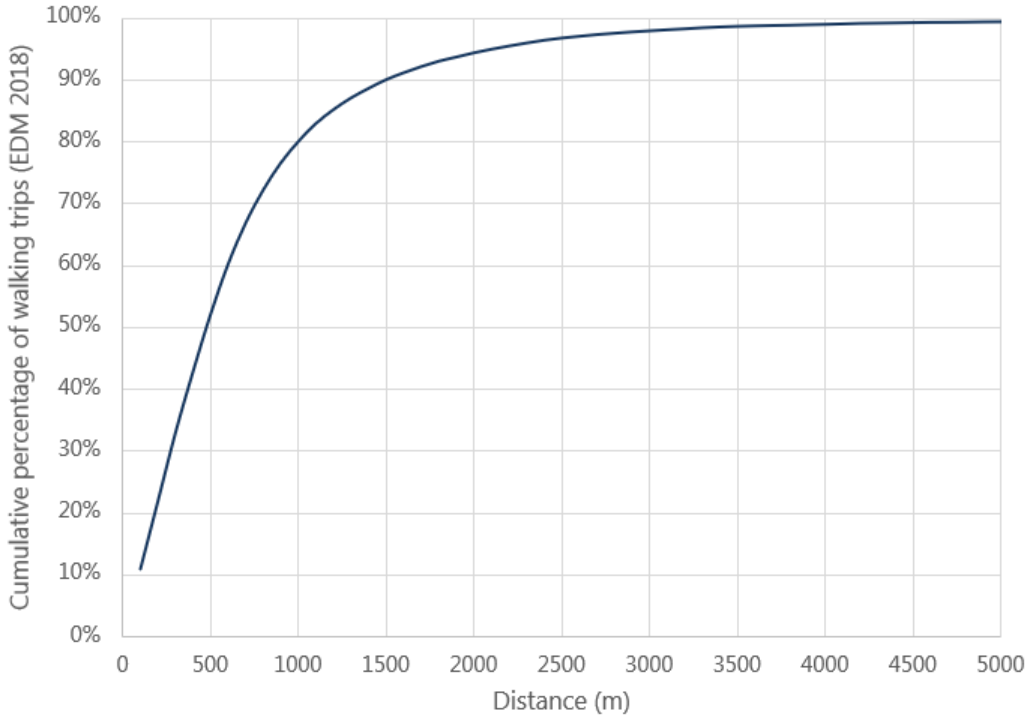


Figure 4.1: Cumulative distance distribution of walking trips according to the household survey

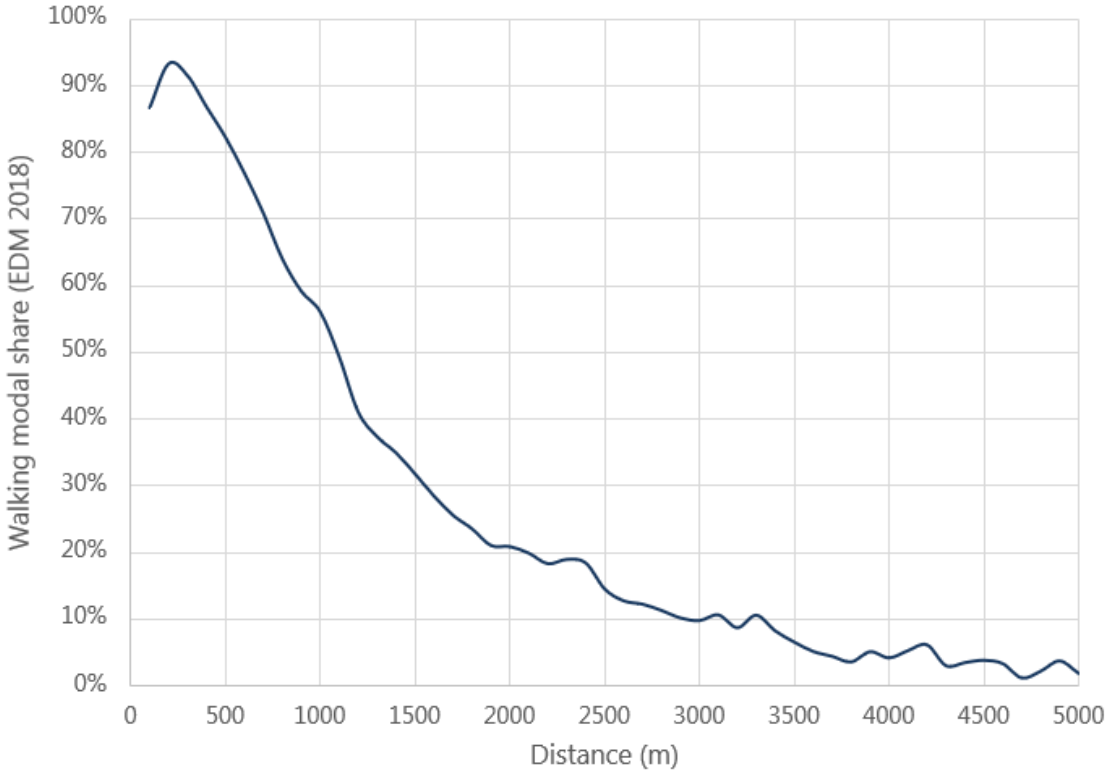


Figure 4.2: Walking modal share depending on trip distances according to the household survey

The distance range with walking modal share between 10% and 90% is approximately 500-3,500 m, so it is interesting to focus on the accuracy of travel times in this range, which is the one where walking competes with other modes. Taking this into account, the travel time results from OSRM were compared against Google Maps for several OD pairs falling in the 0m to 5,000m range.

Tests

To analyse which OSRM walking speed works best, the 4.5 km/h and 5 km/h were defined as an input in OSRM and subsequently different routes were tested and then compared against Google Maps. The table illustrates this comparison showing the different travel times and distances returned for each tested route. It must be pointed out that both OSRM and Google Maps return more than one route for each query, being option 1 the optimal route and option 2 the second-best option.

Table 4.25. Walking speeds tests analysis

	Google Maps (option 1)		Google Maps (option 2)		5km/h (option 1)			5km/h (option 2)			4.5km/h (option 1)			4.5km/h (option 2)		
OD pair	km	t (min)	km	t (min)	km	t (min)	Diff (min)	km	t (min)	Diff (min)	km	t (min)	Diff (min)	km	t (min)	Diff (min)
Puerta de Toledo to Ibiza	4.0	53	4.1	55	3.9	47	6		0		3.8	51	2		0	
Ibiza to Argüelles	3.8	48	3.9	49	3.9	47	1	4.1	49	0	3.9	53	-5	4.1	55	0
Argüelles to 4 Caminos	2.4	32	2.4	33	2.5	31	1	2.5	31	2	2.5	34	-2	2.5	34	2
4 Caminos to Cuzco	2.3	30	2.4	30	2.3	28	2	2.4	30	0	2.3	32	-2	2.4	33	0
Cuzco to Pinar del Rey	4.4	55	4.5	57	3.9	47	8		0		3.9	53	2		0	
San Bernardo to Tribunal	0.7	9	0.8	10	0.6	9	0	0.7	9	1	0.6	8	1	0.7	8	1
Opera to Sol	0.5	7	0.6	8	0.6	7	0	0.6	9	-1	0.6	7	0	0.6	8	-1
Pirámides to Lavapiés	1.3	17	1.3	17	1.2	16	1	1.2	17	0	1.2	14	3	1.2	15	0
Alcalá Centro - Estación	1.1	14	1.2	15	1.1	14	1		0		1.1	14	1		0	
Getafe (Juan de la Cierva) - Centro	1.1	14	1.2	14	1.1	14	0	1.2	15	-1	1.1	14	0	1.2	17	-1

When analysing the routes followed for each query, all searched routes, except Cuzco to Pinar del Rey, returned similar paths in OSRM and Google Maps. For Cuzco to Pinar del Rey, the OSRM returned route had a section on a motorway (highly unlikely when walking, see Figure 4.3). Regarding the speeds, the travel times derived from having a walking speed of 4.5km/h are generally better aligned with the ones from Google Maps.

Comparing option 1 (the optimal option) for each walking speed, it can be observed that the average walking speed of 5 km/h has an average difference of 2 minutes, while the average walking speed of 4.5 km/h has an average difference of 0 minutes. However, when it comes to short trips, which are more likely to be done by foot, 5km/h seems to be more aligned with Google Maps.

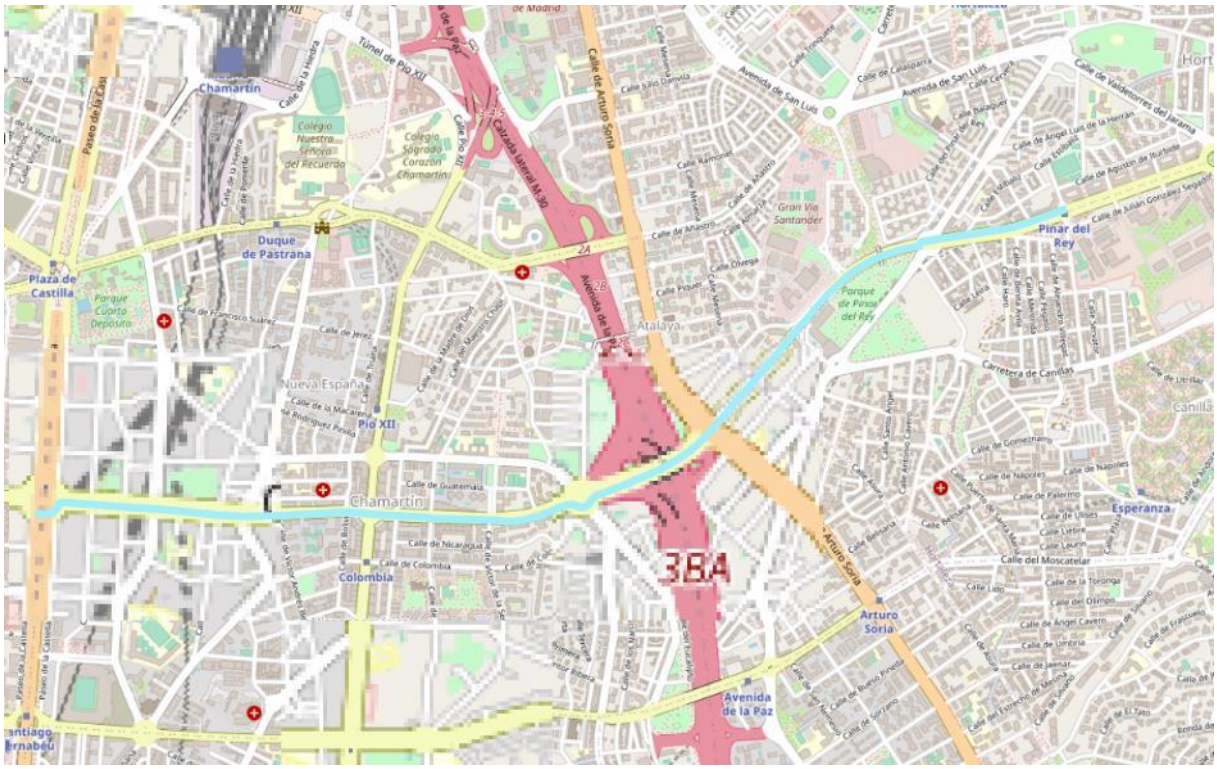


Figure 4.3: Walking route Cuzco-Pinar del Rey through a motorway junction without sidewalks

Conclusions

It can be observed that the walking speed of 4.5km/h returns travel times that are closer to the ones from Google Maps in general, whereas the walking speed of 5 km/h returns travel times that are closed to the ones from Google Maps for shorter trips. Knowing that the majority of walking trips have a short distance, it has been concluded that the best is to set walking speed at the default value of 5 km/h.

Public Transport (Open Trip Planner)

Inputs

The code requesting the route query includes required the following parameters:

- Origin, Destination
- Time of departure (departure_hour)
- Day of departure (departure_date)
- Maximum walking distance (sets the maximum walking distance as a constraint that can be broken if there is no other alternative, i.e., if there are two alternatives but one exceeds the maximum walking distance, only one alternative would be presented, but if there is only one alternative and this one exceeds the maximum walking distance, this one would be presented nevertheless)
- Mode used (mode_used) ("TRANSIT,walk")

In addition, OTP requires General Traffic Feed Specification (GTFS) files which detail the transport supply characteristics. Initially, only metro and urban bus GTFS were used. Later on, Cercanías (which is the urban rail in Spain), light-rail and inter-urban buses were added.

Methodology

Public transport trips are not bounded in distance terms as it is the case for walking trips (Figure 4.4). About 75% of the trips are below 10 km long. A vast majority of the public transport trips above this distance are made by commuter railway or interurban bus (Figure 4.5). The joint modal share of these two modes is significant in any distance range compared to private vehicles, around 25-30% (Figure 4.6). This implies that an accurate description of the public transport supply is required at any distance range.

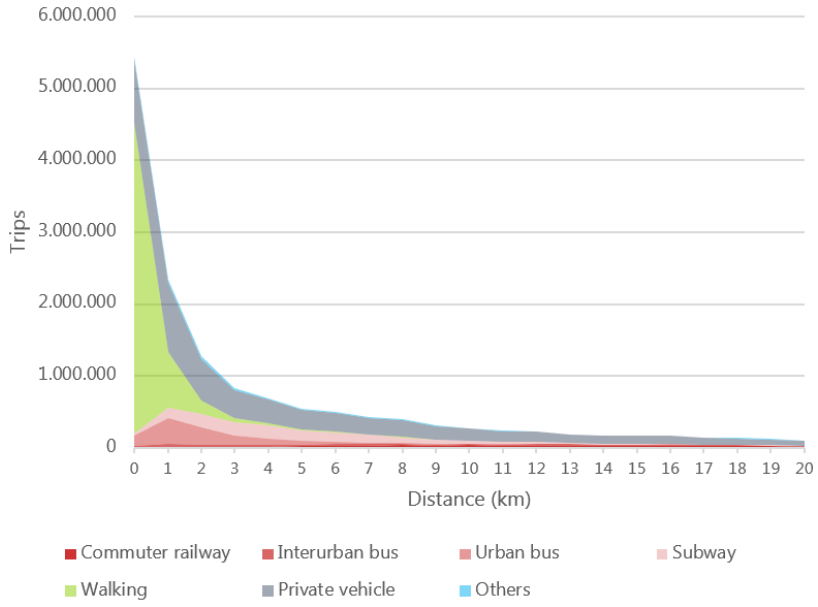


Figure 4.4: Trip distance distribution segmented by mode according to the household survey

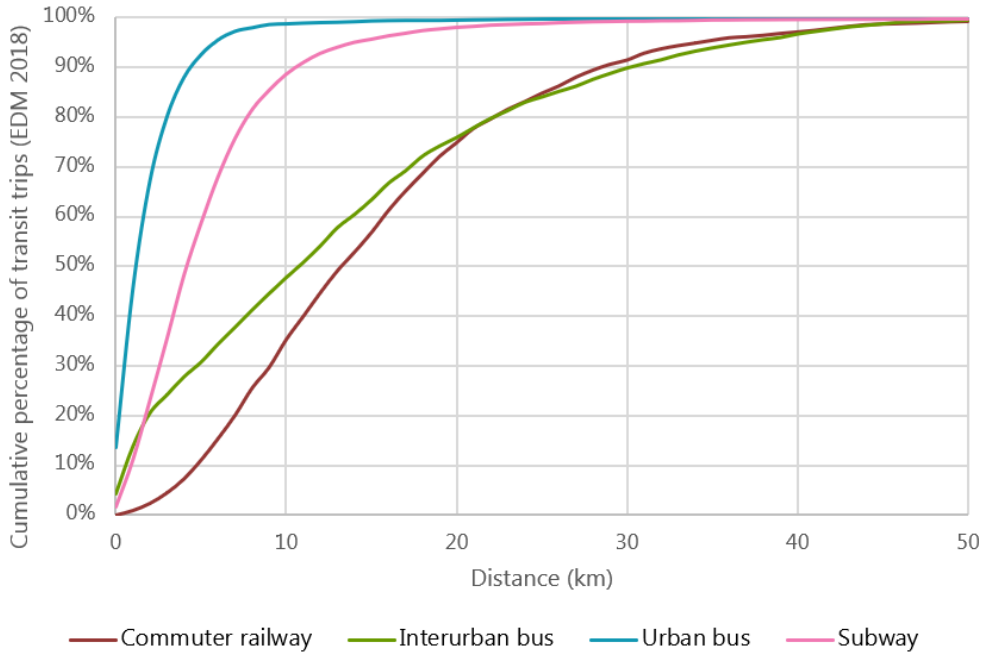


Figure 4.5: Cumulative distance distribution of PT modes trips according to household survey

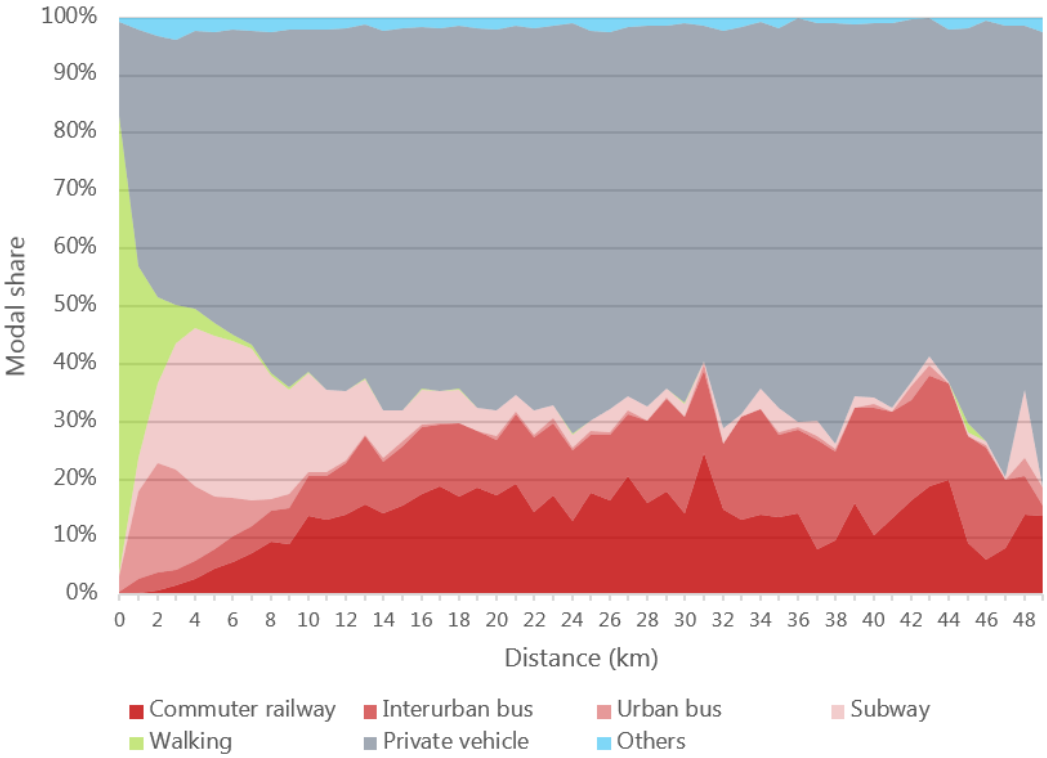


Figure 4.6: Modal share depending on trip distances according to household survey

Following these analyses, it is clear that it is necessary to perform tests over a wide range of OD pairs with different distances. These tests consist in comparing the results from OTP and Google Maps and ensuring that there are no significant differences between them. There are two approaches to choose OD pairs:

- Grid-oriented. This would be based on the establishment of a grid of points (e.g. 1,000m) and retrieving queries from both Google Maps and OTP in order to compute statistics about the differences in travel times between both sources.
- Case-oriented. This would be based on the selection of specific OD pairs that are representative of the public transport demand and observe the differences between the travel times offered by both sources.

Initially the second approach is selected, as it allows us to have a better understanding of the deviation causes and check the performance of OTP under controlled conditions. The grid-oriented approach is reserved for a second test wave to check the overall performance of OTP if required.

The selection of OD pairs under the case-oriented approach leads to two sets of cases:

- OD pairs between points at the target stations or stops, to check the performance of OTP in comparison to Google Maps regardless of the access/egress times (focus on in-vehicle and waiting times).
- OD pairs between points that may be representative of door-to-door trips, to check the performance of OTP in comparison to Google Maps taking into account access/egress times.

For each OD pair, queries are made to Google Maps and OTP services for a set of different periods of the week and day, to see differences in the OTP performance depending on the day of the week and the time of the day.

Table 4.26: OD pairs between stations or stops

Origin	Destination	Motivation
Puerta de Toledo	Ibiza	Check subway travel times estimation
Ibiza	Arguelles	Check subway travel times estimation
Arguelles	4 Caminos	Check subway travel times estimation
4 Caminos	Cuzco	Check subway travel times estimation
Cuzco	Pinar del Rey	Check subway travel times estimation
San Bernardo	Tribunal	Check subway travel times estimation
Opera	Sol	Check subway travel times estimation
Piramides	Lavapiés	Check subway travel times estimation
Banco de España	Islas Filipinas	Check subway travel times estimation
Banco de España	Guzman el Bueno	Check subway travel times estimation
Banco de España	Moncloa	Check subway travel times estimation
Ruben Darío (Castellana)	Plaza España	Check subway travel times estimation
Atocha	Nuevos Ministerios	Check commuter railway travel times estimation (direct link)

Origin	Destination	Motivation
Santa Eugenia	Aravaca	Check commuter railway travel times estimation (direct link)
Villaverde Alto	Aravaca	Check commuter railway travel times estimation (direct link)
Atocha	Aravaca	Check commuter railway travel times estimation (direct link)
Santa Eugenia	Aravaca	Check commuter railway travel times estimation (non-direct link)
Villaverde Alto	Aravaca	Check commuter railway travel times estimation (non-direct link)
Santa Eugenia	Nuevos Ministerios	Check commuter railway travel times estimation (non-direct link)
Villaverde Alto	Nuevos Ministerios	Check commuter railway travel times estimation (non-direct link)
Alcalá	Avenida America	Check interurban trips (direct link)
Principe Pío	Boadilla	Check interurban trips (direct link)
Principe Pío	Pelayos de la Presa	Check interurban trips (direct link)
Chamartín	3 Cantos	Check interurban trips (direct link)
A-1 corridor		
Stop “Marqués de Valdavia - Av.España” (Alcobendas)	Stop “Intercambiador Plaza de Castilla” (Madrid)	Check interurban bus travel times estimation (direct link) 1 leg observed in EDM (H3275061, I1, T1, L1)
Stop “Avda Madrid-Residencial Guadalix” (San Agustín de Guadalix)	Stop “Bulevar Salvador Allende-Soria” (Alcobendas)	Check interurban bus travel times estimation (direct link) 1 leg observed in EDM (H3552427, I1, T1, L1)
Station “Alcobendas-San Sebastián de los Reyes”	Station “Sol”	Check commuter railway travel times estimation (direct link) 20 legs observed in EDM (e.g., H1417637, I1, T2, L1)
A-2 corridor		
Stop “Vía Complutense - Brihuega” (Alcalá de Henares)	Stop “Intercambiador Avenida de América” (Madrid)	Check interurban bus travel times estimation (direct link) 6 legs observed in EDM (e.g., H3344388, I1, T3, L1) Obs: apparently not available in GMaps, to be checked using Citymapper

Origin	Destination	Motivation
Stop “Daganzo de Arriba-Gta.Alcalá” (Daganzo de Arriba)	Stop “Vía Complutense - Barrio Ledesma” (Alcalá de Henares)	Check interurban bus travel times estimation (direct link) 1 leg observed in EDM (H10731, I3, T2, L1) Obs: apparently not available in GMaps, to be checked using Citymapper
Station “Alcalá de Henares”	Station “Vallecas”	Check commuter railway travel times estimation (direct link) 8 legs observed in EDM (e.g., H2000432, I2, T1, L2)
A-3 corridor		
Stop “Ronda de Atocha-Estación de Autobuses” (Madrid)	Stop “Av.valencia - pza.progreso” (Arganda del Rey)	Check interurban bus travel times estimation (direct link) 1 leg observed in EDM (H3477240, I1, T1, L3)
Stop “Av.Mediterráneo-Conde Casal” (Madrid)	Stop “Av.Pablo Iglesias - Jovellanos” (Arganda del Rey)	Check interurban bus travel times estimation (direct link) 2 legs observed in EDM (e.g., H3054318, I1, T2, L3) Obs: apparently not available in GMaps, to be checked using Citymapper
A-4 / A-42 corridor		
Stop “Calle Madrid-Universidad” (Getafe)	Stop “Intercambiador Plaza Elíptica” (Madrid)	Check interurban bus travel times estimation (direct link) 5 legs observed in EDM (e.g., H1200813, I2, T2, L1)
Stop “Parla-Vicente Aleixandre” (Valdemoro)	Stop “Legazpi” (Madrid)	Check interurban bus travel times estimation (direct link) 2 legs observed in EDM (e.g., H2431129, I1, T1, L1) Obs: apparently not available in Google Maps, to be checked using Citymapper
Stop “Luis Sauquillo-Grecia” (Fuenlabrada)	Stop “Santiago Ramón y Cajal - Severo Ochoa” (Humanes de Madrid)	Check interurban bus travel times estimation (direct link) 5 legs observed in EDM (e.g., H550032, I1, T2, L1) Obs: apparently not available in Google Maps, to be checked using Citymapper
Station “Parla”	Station “Cantoblanco Universidad”	Check commuter railway travel times estimation (direct link) 5 legs observed in EDM (e.g., H2142309, I2, T1, L1)
A-5 corridor		
Stop “Av. Carlos V - Liceo Villafontana” (Móstoles)	Stop “Príncipe Pío” (Madrid)	Check interurban bus travel times estimation (direct link) 12 legs observed in EDM (e.g., H2044461, I3, T1, L1)

Origin	Destination	Motivation
Stop "Crta.Fuenlabrada--Serranillos" (Arroyomolinos)	Stop "Príncipe Pío" (Madrid)	Check interurban bus travel times estimation (direct link) 4 legs observed in EDM (e.g., H230649, I6, T1, L1)
Station "Alcorcón"	Station "Atocha"	Check commuter railway travel times estimation (direct link) 33 legs observed in EDM (e.g., H36531, I4, T2, L1)
A-6 corridor		
Stop "J.Rodrigo - Hospital Puerta de Hierro" (Majadahonda)	Stop "Intercambiador de Moncloa" (Madrid)	Check interurban bus travel times estimation (direct link) 11 legs observed in EDM (e.g., H1981625, I3, T1, L1)
Stop "Alfonso Senra-Pza.Mayor" (Guadarrama)	Stop "Intercambiador de Moncloa" (Madrid)	Check interurban bus travel times estimation (direct link) 9 legs observed in EDM (e.g., H541135, I3, T1, L1)
Station "Collado Villalba"	Station "Aravaca"	Check commuter railway travel times estimation (direct link) 2 legs observed in EDM (e.g., H318498, I2, T1, L1)
Additional OD pairs		
Buitrago de Lozoya	Aranjuez	Check map boundaries
Yuncos	Aranjuez	Check map boundaries
Yuncos	Pelayos de la Presa	Check map boundaries
Buitrago de Lozoya	Pelayos de la Presa	Check map boundaries
Buitrago de Lozoya	Miraflores de la Sierra	Check map boundaries
Alpedrete	Guadarrama	Check map boundaries
Parla	Pinto	Check map boundaries
Parla	Aranjuez	Check map boundaries
Pelayos de la Presa	Navas del Rey	Check map boundaries

Tests

Sensitivity to maximum walking distance parameter. When running OTP, the origin, destination, time of departure and maximum walking distance are inputs. The following table illustrates the route summary for different OD pairs. In this case, the default input maximum walking distance was 500 m. At first glance, it can be observed that travel times tend to be higher. This was found to be due to the constraint in maximum walking distance. This variable specifies that if any route or routes can be completed without walking over a certain distance, the algorithm will return the best route among the ones that meet these criteria. If no routes are found meeting these criteria, the constraint will be broken anyway until a route is found. Hence, having a small maximum walking distance can push the algorithm into finding suboptimal routes. This motivated increasing the value of this parameter in the subsequent tests.

Table 4.27: Saturday 21/11/2020 at 2.02 pm table (walking distance = 500m - default)

Saturday										
		oogle Maps (option		oogle Maps (option		oogle Maps (option		500m (option1)	500m (option2)	500m (option3)
	OD pair	mode	t (min)	mode	t (min)	mode	t (min)	t (diff)	t (diff)	t (diff)
0	Puerta de Toledo to Ibiza	subway	29	subway	30	subway	36	16	21	25
1	Ibiza to Argüelles	subway	26	subway	30	subway	37	11		
2	Argüelles to 4 Caminos	subway	15	subway	19	bus	21	1	9	7
3	4 Caminos to Cuzco	subway	17	subway	22	subway	23	20	14	17
4	Cuzco to Pinar del Rey	subway	24	subway	28	combined	33	7	25	21
5	San Bernardo to Tribunal	**walk**	**9**					**0**	13	16
6	Opera to Sol	**walk**	**6**	subway	7	subway	12	**0**	3	0
7	Piramides to Lavapiés	**walk**	**17**	bus	20	subway	19	**5**	3	4
8	Banco de Espana to Islas Filipinas	subway	23	subway	24	bus	39	23	23	7
9	Banco de Espana to Guzman el Bueno	subway	27	subway	30	subway	32	22	20	18
10	Banco de Espana to Moncloa	subway	20	bus	32	combined	29	6	4	8
11	Ruben Darío / Castellana to Plaza España	subway	13	subway	13	subway	13	8	18	

Table 4.28: Monday 23/11/2020 at 2.02 pm table (walking distance = 500m - default)

Monday										
		oogle Maps (option		oogle Maps (option		oogle Maps (option		500m (option1)	500m (option2)	500m (option3)
	OD pair	mode	t (min)	mode	t (min)	mode	t (min)	t (diff)	t (diff)	t (diff)
0	Puerta de Toledo to Ibiza	subway	28	subway	29	bus	40	17	15	
1	Ibiza to Argüelles	subway	24	subway	27	subway	26	10	16	19
2	Argüelles to 4 Caminos	subway	14	subway	17	subway	18	1	11	11
3	4 Caminos to Cuzco	subway	14	subway	19	subway	19	18		
4	Cuzco to Pinar del Rey	subway	21	subway	27	subway	26	6		
5	San Bernardo to Tribunal	**walk**	**9**	subway	10	subway	12	0	1	
6	Opera to Sol	**walk**	**6**	subway	6	subway	11	0	5	
7	Piramides to Lavapiés	**walk**	**17**	bus	17	subway	17			
8	Banco de Espana to Islas Filipinas	subway	21	subway	24	bus	36	26	23	12
9	Banco de Espana to Guzman el Bueno	subway	26	subway	28	bus	40	17	15	
10	Banco de Espana to Moncloa	subway	18	subway	23	bus	30	5	13	
11	Ruben Darío / Castellana to Plaza España	subway	12	bus	26	walk	32	7	1	

In a second iteration, the same routes were tested with a **1,000m maximum walking distance**. This reduced the differences between OTP estimation and Google Maps estimation for 10 out of the 12 cases tested.

Table 4.29: Saturday 21.11.2020 table (walking distance = 1,000m - default)

Saturday												
		oogle Maps (option		oogle Maps (option		oogle Maps (option		1000m (option1)		1000m (option2)		1000m (option3)	
	OD pair	mode	t (min)	mode	t (min)	mode	t (min)	mode	t (diff)	mode	t (diff)	mode	t (diff)
0	Puerta de Toledo to Ibiza	subway	29	subway	30	subway	36	bus	12	bus	11	bus	9
1	Ibiza to Argüelles	subway	26	subway	30	subway	37	**subway**	-1	subway	**8**		
2	Argüelles to 4 Caminos	subway	15	subway	19	bus	21	**subway**	1	subway	**6**	bus	**7**
3	4 Caminos to Cuzco	subway	17	subway	22	subway	23	subway	11	bus	15	bus	16
4	Cuzco to Pinar del Rey	subway	24	subway	28	combined	33	subway	18	bus	**28**		
5	San Bernardo to Tribunal	**walk**	**9**					**walk**	**0**	subway	**13**	subway	**15**
6	Opera to Sol	**walk**	**6**	subway	7	subway	12	**walk**	**0**	subway	**3**	subway	**0**
7	Piramides to Lavapiés	**walk**	**17**	bus	20	subway	19	**walk**	**0**	subway	**3**	subway	**5**
8	Banco de Espana to Islas Filipinas	subway	23	subway	24	bus	39	bus	21	bus	22	bus	7
9	Banco de Espana to Guzman el Bueno	subway	27	subway	30	subway	32	**subway**	4	bus	20	bus	18
10	Banco de Espana to Moncloa	subway	20	bus	32	combined	29	**subway**	2	bus	4	bus	8
11	Ruben Darío / Castellana to Plaza España	subway	13	subway	13	subway	13	**subway**	7	subway	10	bus	18

Table 4.30: Monday 23.11.2020 table (walking distance = 1000m - default)

Monday													
		oogle Maps (option		oogle Maps (option		oogle Maps (option		1000m (option1)		1000m (option2)		1000m (option3)	
	OD pair	mode	t (min)	mode	t (min)	mode	t (min)	mode	t (diff)	mode	t (diff)	mode	t (diff)
0	Puerta de Toledo to Ibiza	subway	28	subway	29	bus	40	bus	15				
1	Ibiza to Argüelles	subway	24	subway	27	subway	26	**subway**	0	subway	9	bus	17
2	Argüelles to 4 Caminos	subway	14	subway	17	subway	18	bus	1	bus	11	bus	11
3	4 Caminos to Cuzco	subway	14	subway	19	subway	19	bus	16	bus	15		
4	Cuzco to Pinar del Rey	subway	21	subway	27	subway	26	subway	16				
5	San Bernardo to Tribunal	**walk**	9	subway	10	subway	12	**walk**	0	subway	1	subway	1
6	Opera to Sol	**walk**	6	subway	6	subway	11	**walk**	0	subway	5		
7	Piramides to Lavapiés	**walk**	17	bus	17	subway	17	**walk**	0				
8	Banco de Espana to Islas Filipinas	subway	21	subway	24	bus	36						
9	Banco de Espana to Guzman el Bueno	subway	26	subway	28	bus	40	**subway**	5	bus	15	bus	3
10	Banco de Espana to Moncloa	subway	18	subway	23	bus	30	**subway**	2	bus	13		
11	Ruben Darío / Castellana to Plaza España	subway	12	bus	26	walk	32	**subway**	6	subway	-4	bus	-4

In subsequent queries, OTP has failed to return a route. To tackle this issue, pushing the maximum walking distance even higher has been an effective measure.

Adjustment of headway values. After adjusting the maximum walking distance, a closer look to the actual routes has been taken. For the figures illustrated below, it can be observed that the selected modes do not match to the ones observed in Google Maps. In addition, unusually cumbersome routes by bus were calculated instead of the easy subway routes. When analysing the waiting times, the values seemed unusually high. This led to investigating whether there was a problem with subway frequencies. Checking the subway headways during the day (AM, IP and PM periods), they spanned from 3.5 to 10.5 minutes, which does seem reasonable. The question is how is OPT calculating the waiting time for a route. On a thorough analysis, checking the end time of each leg and the starting time of the subsequent leg, it has been observed that the waiting times correspond to the service headways. However, this corresponds to the maximum waiting time and not the expected waiting time. Hence, in an attempt to correct this, the headways have been divided by two. After this correction, travel times seem to be much more aligned with the ones from Google Maps and the same happens with the selected mode. The results are as follows:

Table 4.31: Saturday 21.11.2020 at 2.02 pm table

Saturday - 2.02pm													
		oogle Maps (option		oogle Maps (option		oogle Maps (option		1000m (option1)		1000m (option2)		1000m (option3)	
	OD pair	mode	t (min)	mode	t (min)	mode	t (min)	mode	t (diff)	mode	t (diff)	mode	t (diff)
0	Puerta de Toledo to Ibiza	subway	29	subway	30	subway	36	bus	8	bus	7	bus	5
1	Ibiza to Argüelles	subway	26	subway	30	subway	37	**subway**	-4	subway	2	bus	6
2	Argüelles to 4 Caminos	subway	15	subway	19	bus	21	**subway**	-2	subway	3	bus	3
3	4 Caminos to Cuzco	subway	17	subway	22	subway	23	**subway**	7	subway	2	bus	8
4	Cuzco to Pinar del Rey	subway	24	subway	28	combined	33	**subway**	1	subway	7		
5	San Bernardo to Tribunal	walk	9					**walk**	-1	subway	9	subway	12
6	Opera to Sol	walk	6	subway	7	subway	12	**walk**	0	subway	0	subway	-3
7	Piramides to Lavapiés	walk	17	bus	20	subway	19	bus	0	**walk**	-3	bus	-2
8	Banco de Espana to Islas Filipinas	subway	23	subway	24	bus	39	**subway**	7	bus	15		
9	Banco de Espana to Guzman el Bueno	subway	27	subway	30	subway	32	**subway**	2	subway	1	subway	0
10	Banco de Espana to Moncloa	subway	20	bus	32	combined	29	**subway**	-1	bus	-1	bus	2
11	Ruben Darío / Castellana to Plaza España	subway	13	subway	13	subway	13	**subway**	4	subway	7	subway	14

Table 4.32: Monday 23.11.2020 table

	Monday - 2.02pm												
		oogle Maps (option		oogle Maps (option		ogle Maps (option		1000m (option1)		1000m (option2)		1000m (option3)	
	OD pair	mode	t (min)	mode	t (min)	mode	t (min)	mode	t (diff)	mode	t (diff)	mode	t (diff)
0	Puerta de Toledo to Ibiza	subway	28	subway	29	bus	40	**subway**		subway		bus	
1	Ibiza to Argüelles	subway	24	subway	27	subway	26	**subway**	-3	subway	3	bus	12
2	Argüelles to 4 Caminos	subway	14	subway	17	subway	18	**subway**	-1	subway	2	bus	7
3	4 Caminos to Cuzco	subway	14	subway	19	subway	19						
4	Cuzco to Pinar del Rey	subway	21	subway	27	subway	26	**subway**	2	subway	5		
5	San Bernardo to Tribunal	walk	9	subway	10	subway	12	**walk**	-1	subway	-2	subway	-1
6	Opera to Sol	walk	6	subway	6	subway	11	**walk**	0	subway	1	subway	-2
7	Piramides to Lavapiés	walk	17	bus	17	subway	17	**walk**	0				
8	Banco de Espana to Islas Filipinas	subway	21	subway	24	bus	36						
9	Banco de Espana to Guzman el Bueno	subway	26	subway	28	bus	40	bus	13	bus	11	subway	-11
10	Banco de Espana to Moncloa	subway	18	subway	23	bus	30	**subway**	0				
11	Ruben Darío / Castellana to Plaza España	subway	12	bus	26	walk	32	**subway**	4	subway	-7	bus	-9

Time periods to be considered. As explained above, OTP fails for some given OD pairs or times of the day. On top of modifying the maximum walking distance, whenever OTP fails to find a route for a given time, an algorithm to test "similar" times has been put in place to search similar queries until OTP finds a valid alternative. The idea behind running OTP is to find travel costs for a specific OD pair. Given that costs change throughout the day, it is believed that having different costs for different parts of the day is a reasonable approach. However, what is the optimal time segmentation? On a first iteration, it is believed that segmenting into AM (07:00-10:00), IP (10:00-16:30), PM (16:30-19:30) and OP (19:30-07:00) is good enough. Consequently, if OPT fails to provide the costs for a given time, it will iterate within the time period until it finds a valid alternative.

Addition of all public transport modes. In order to capture all modes included in public transport, light-rail (tramway), interurban buses and Madrid´s Cercanías have been included in OTP and routes in which these modes are likely to be chosen have been tested. In addition, the maximum walking distance has been pushed to 10,000 metres to avoid any potential errors.

Table 4.33: Monday 23.11.2020 table with more test and modes

	Monday - 2.02pm												
		oogle Maps (option		oogle Maps (option		ogle Maps (option		10,000m (option1)		10,000m (option2)		10,000m (option3)	
	OD pair	mode	t (min)	mode	t (min)	mode	t (min)	mode	t (diff)	mode	t (diff)	mode	t (diff)
Rail	**Atocha to Corteinglés (Nuevos Ministerios)**	rail	12	rail	14	rail	12	**rail**	13	**rail**	12	**rail**	18
Rail	Atocha to Aravaca	all	65	all	64	all	70	bus	11	bus	12	bus	
Rail	Santa Eugenia to Aravaca	all	76	all	81	all	84	**rail**	18	rail+bus	15	rail	26
Rail	Villaverde to Aravaca	all	52	combined	52	combined	55	rail+bus	26	rail+bus	27	rail+bus	34
Rail	Santa Eugenia to Corteinglés (Nuevos Ministerios)	rail	25	rail	25	rail	25	**rail**	0	**rail**	13	**rail**	30
Rail	Villaverde to Corteinglés (Nuevos Ministerios)	rail	30	rail	30	rail	34	**rail**	-4	**rail**	0	**rail**	7
Interurban	Alcalá to Avenida America	rail+subw	88	rail+subw	91	rail+inter	91	**inter bus**	107	inter bus	125	inter bus	179
Interurban	Principe Pío to Boadilla	inter bus	54	subway+t	51	inter bus	54	**inter bus**	3	inter bus	36	**inter bus**	38
Interurban	Principe Pío to Pelayos de la Presa	inter bus	78	inter bus	78			**inter bus**	-1	inter bus	-1	inter bus	77
Interurban	Chamartín to 3 Cantos	inter bus	40	inter bus	45	inter bus	45	rail+bus	14	**inter bus**	10	inter bus	14
A1 corridor	Marqués Valvia to Parking Plaza Castilla	inter bus	40	inter bus	51	rail	37	inter bus	5	**inter bus**	-1	inter bus	18
A1 corridor	Av Madrid to Calle de Soria	inter bus	71	inter bus	85	inter bus	96						
A1 corridor	Alcobendas to Sol	rail	40	rail	63	rail	55	**rail**	-1	rail	-1	rail	24
A2 corridor	Alcala to Av America	inter bus	43	inter bus	53	inter bus	63	rail+subwa	31	rail+subwa	47	rail+subwa	40
A2 corridor	Daganzo (Alcala) to Ledesma (Alcala)	inter bus	37	inter bus	77	inter bus	46	walk	132	inter bus	99	inter bus	132
A2 corridor	Alcala to Vallecas	rail	35	rail	48	rail	64	**rail**	0	rail	0	rail	0
A3 corridor	La Poveda to Atocha (bus)	inter bus	89	rail+subw	72	inter bus	131	**inter bus**	17	inter bus	34	inter bus	-24
A3 corridor	Av Mediterraneo to Rivas	inter bus	44	subway	66	combined	83	**inter bus**	10	inter bus	-7	subway	-16
A4 corridor	University Carlos III to Intercambiador Plaza Elíptica	inter bus	49	inter bus	55	inter bus	64	**inter bus**	-12	inter bus	-5		
A4 corridor	Valdemoro to Legazpi	combined	110	combined	130	combined	245	rail	-12	rail	-13	rail	-108
A4 corridor	Fuenlabrada to Humanes	rail	29	rail	41	rail	53	**rail**	1	inter bus	-2	walk	-13
A4 corridor	Parla to Cantoblanco Universidad	rail	54	rail	72	rail	73	**rail**	0	rail	0	rail	20
A5 corridor	Mostoles to Principe Pío	inter bus	45	inter bus	50	inter bus	52	**inter bus**	-12	inter bus	-17	inter bus	-12
A5 corridor	Arroyomolinos to Principe Pío	inter bus	36	inter bus	54	inter bus	57	**inter bus**	-2	inter bus	-1	inter bus	-4
A5 corridor	Alcorcón to Atocha	rail	25	rail	31	rail	37	**rail**	4	rail	4	rail	4
A6 corridor	Mahadaonda to Moncloa	inter bus	58	inter bus	54	inter bus	73	**inter bus**	-9	inter bus	-5	inter bus	-15
A6 corridor	Guadarrama to Moncloa	inter bus	53	inter bus	63	inter bus	73	**inter bus**	61	inter bus	61	inter bus	151
A6 corridor	Villalba to Arabaca	rail	46	rail	107	all	91	**rail**	3	inter bus	-26	rail	3
missing	Chamartín to Escorial												
missing	Atocha to Escorial												

Failures of the OTP service in certain areas. For some locations, OTP has not been able to find a route. However, the number of unreturned routes is not significant and these routes have been omitted for subsequent analysis.

Conclusions

It has been observed that the OTP algorithm is less faulty (i.e., it retrieves valid public transport options for more cases) with higher maximum walking distances. Consequently, a relatively high walking distance is recommended.

In addition, it has been observed that the OTP algorithm does not calculate the expected waiting time, but instead it uses the maximum waiting time. When it comes to frequency-based services (vs. scheduled-based services), this value is twice as big as the expected one. Consequently, the input frequencies have been divided by two, so that the algorithm calculates an average waiting time that is more representative of the real situation.

In order to limit the number of queries to a bounded number, the day has been divided into different time periods (AM, IP, PM and OP). The transport supply is considered to be constant throughout the time period, which is not far from the truth, so that only "one" query has to be launched per OD pair, time period and day type. However, OTP sometimes returns empty routes, depending on the centroid location and time period. To minimise the number of empty routes, an algorithm has been put in place to launch several queries along the time period and same day types, if needed, until a route is found. This, together with the maximum walking distance parameter tuning explained above, reduces the number of empty routes to a bare minimum.

The OTP results have been compared with Google Maps. On a general note, it can be seen that better results are found in the city centre.

4.2.2. Adding cost to trip characterisation

Parking Categorical Variable

Given that no detailed data on parking availability has been found, a categorical variable has been used for charactering the parking availability.

Inputs

The data inputs are the following

- Calculated routes from OSRM (if the route is by car or by foot) or OTP (if the route is by PT).
- Station and stops to tariff zone correspondence
- Grid to parking zoning correspondence

Methodology

Under the assumption that parking is scarcer as one gets closer to the city centre, a categorical variable has been defined after an iterative testing process. The initial parking categorical variable that has been tested is based on the Madrid Regional Transport Authority tariff zoning (find zone categorisation in the Figure 4.7):

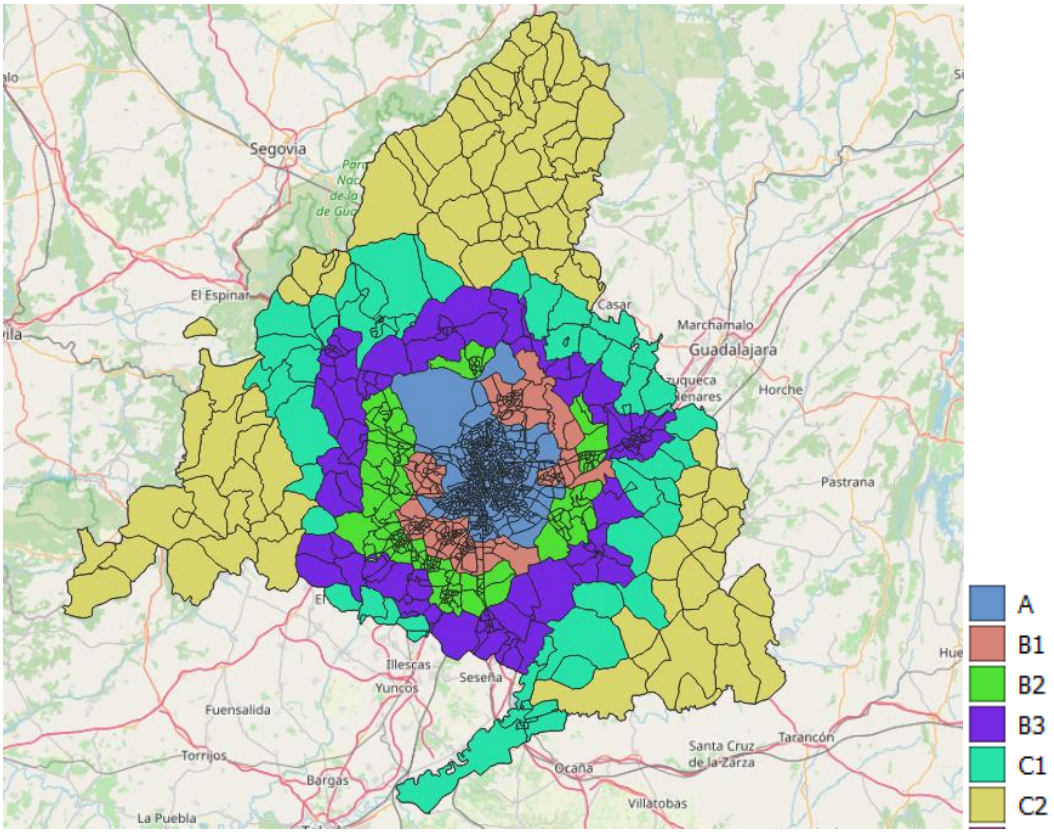


Figure 4.7: Tariff Zoning

This assumption has been tested against the results of the survey in terms of parking availability at the destination for each trip observed. The results from the survey with the given zoning are illustrated below (see Figure 4.8 and Table 4.34):

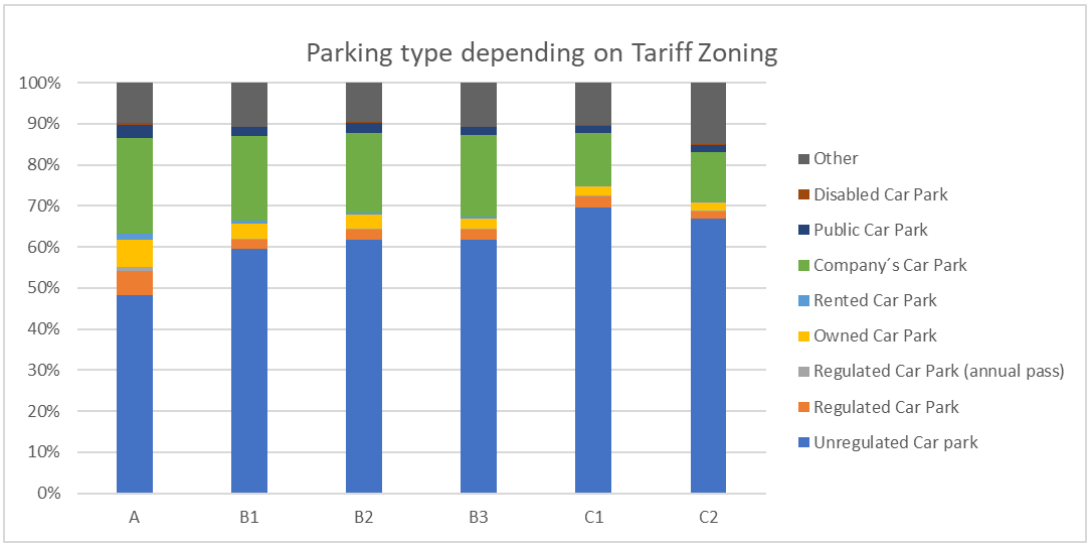


Figure 4.8: Parking results with Tariff Zoning

Table 4.34: Parking results with Tariff Zoning

	A	B1	B2	B3	C1	C2
Unregulated Car park	48%	59%	62%	62%	**69%**	**67%**
Regulated Car Park	6%	2%	2%	2%	3%	2%
Regulated Car Park (annual pass)	1%	0%	0%	0%	0%	0%
Owned Car Park	7%	4%	3%	3%	2%	2%
Rented Car Park	2%	1%	1%	0%	0%	0%
Company´s Car Park	23%	**21%**	**19%**	**20%**	13%	12%
Public Car Park	3%	2%	2%	2%	2%	2%
Disabked Car Park	0%	0%	0%	0%	0%	0%
Other	10%	11%	9%	11%	10%	15%
sample size	28741	12360	14004	9851	4433	2276

From this initial iteration, it is observed that B1, B2 and B3 zones have similar patterns in terms of parking availability. Similarly, C1 and C2 have similar patterns. The parking availability situation varies from one area to another inside the A zone. In addition, part of zone A is restricted to car access (Madrid Central area). To check this hypothesis, in a subsequent iteration, zone A is divided into Madrid's centric district ("Madrid Central"), whatever is inside the M30 ring road (inner city). and whatever is outside M30. Additionally, B1-B2-B3 and C1-C2 are aggregated (see Figure 4.9 below). This leaves 5 zones: 0 (district Centro), 1 (rest of inner city within M30 ring road), 2 (rest of zone A), 3 (zones B) and 4 (zones C).

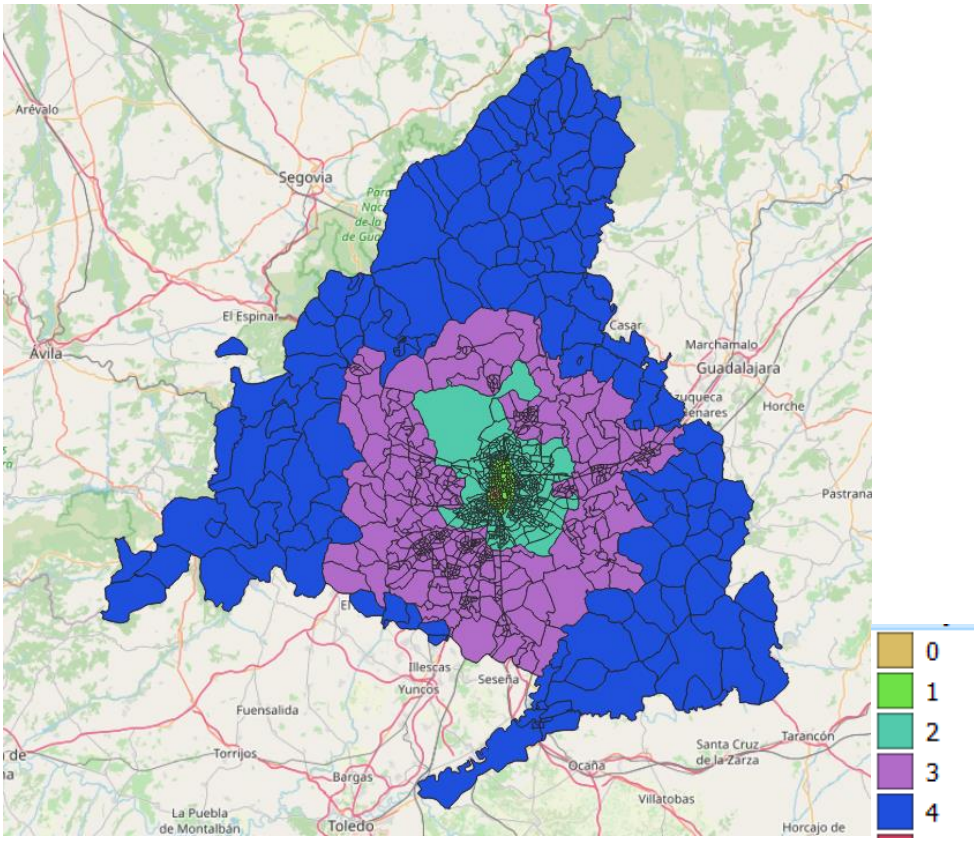


Figure 4.9: Proposed zoning

As it can be observed in Figure 4.10 and Table 4.35, there are no clear differences between zone 0 and zone 1. However, the other zones do show clear differentiated patterns.

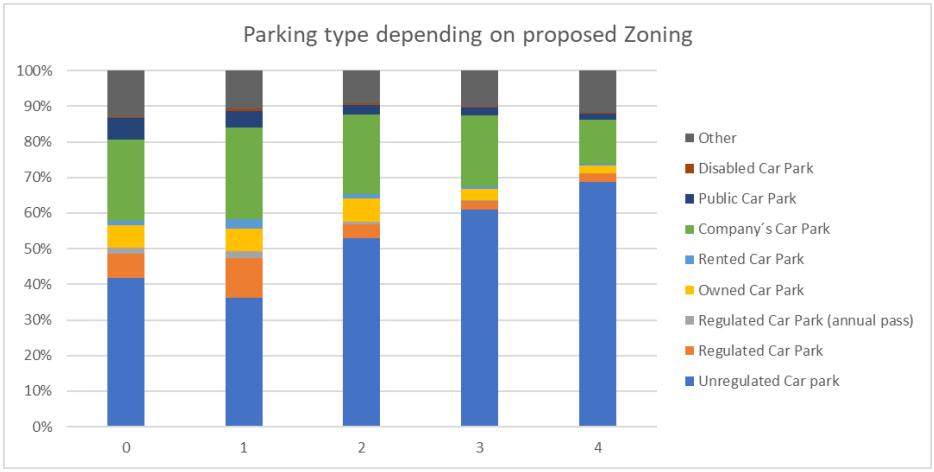


Figure 4.10: Parking results with proposed zoning

Table 4.35: Parking results with proposed zoning

	0	1	2	3	4
Unregulated Car park	42%	36%	53%	61%	69%
Regulated Car Park	7%	11%	4%	2%	3%
Regulated Car Park (annual pass)	2%	2%	1%	0%	0%
Owned Car Park	6%	6%	7%	3%	2%
Rented Car Park	1%	3%	1%	1%	0%
Company´s Car Park	23%	26%	22%	20%	12%
Public Car Park	6%	5%	3%	2%	2%
Disabked Car Park	0%	1%	0%	0%	0%
Other	13%	11%	9%	10%	12%
sample size	1201	7231	20809	36215	6709

Conclusions

From this analysis, it can be extracted that using 4 different zones to differentiate the different destination types and the availability of parking at these zones is reasonable. The different zones correspond to the following:

- whatever is inside the Madrid's M30 (zone 0);
- whatever is outside the Madrid's M30 and inside the tariff zone A (zone1),
- whatever is inside the tariff zones B1, B2 and B3 (zone2); and
- whatever is inside tariff zones C1 and C2 (zone3).

Road Cost

Whenever segmenting the demand by mode, the costs of an option have to be evaluated. These costs can include time costs as well as monetary costs. The results from OSRM and OTP (from urban routes extraction) do not include any monetary costs. In order to assign an operation cost to the trips made by car, the literature has been consulted. A wide range of prices have been found in the literature. Consequently, different prices per km will be tested depending on the project's needs and the most sensible price will be chosen accordingly.

Public Transport Cost

When it comes to public transport, the monetary cost is the ticket price.

Inputs

The main inputs to compute PT cost are the following:

- The route characterisation, which is derived from the urban route extraction development and includes the travel time, distance, trip legs and stops/stations
- Correspondence between the different mode stops and their correspondent tariff zone
- A correspondence between the different tariff zone combinations and the corresponding prices, differentiated between single and season tickets and in the latter case between reduced and not reduced, which is based on age.

Methodology

To better understand PT costs, a thorough analysis of the household travel survey has been carried out. The analysis focused on identifying the distribution of ticket types across PT users, in order to identify the errors associated with different hypotheses regarding the monetary costs of PT options.

Madrid's public transport card (TTP) is the most commonly owned card among PT users, whereas non-frequent PT users have a very even distributed ownership between TTP, multi-ticket card and no card at all. Similarly, a majority of PT users hold a season ticket (see Figure 4.11). A greater trip frequency implies a higher likelihood of having a season ticket.

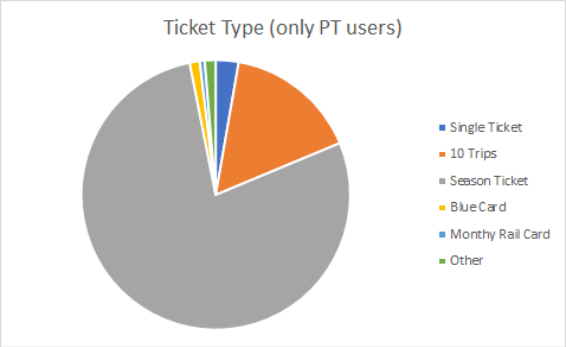


Figure 4.11: Ticket type for PT users

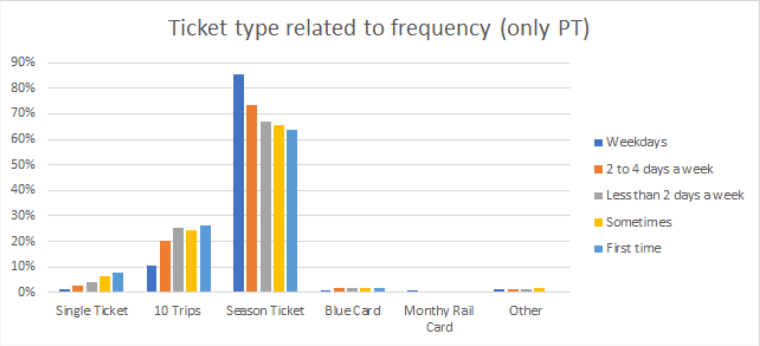


Figure 4.12: Ticket types distribution for different PT usage patterns

The survey also shows that the availability of a TTP card is more common among younger and older segments, probably related to the lower price paid by both groups.

An analysis on the ticket price based on the ticket type has been carried out. Figure 4.13 compares, for different OD pairs, the costs of a single ticket, a 10-trip ticket and a monthly ticket (assuming 44 trips per month, which is derived from 2 trips per day for the usually assumed 22 workdays a month).

Route Test						
Origin	Destination			Single	10 Tickets	Monthly
Parla	**Aranjuez**					
B2-B1	RAIL	B2	B1		1.70	1.00	
B1-B1	BUS	B1	B1		1.30	0.85	
B2-C1	RAIL	B2	C1		1.85	1.37	
B2	**C1**		**3**		**4.85**	**3.22**	**1.24**
B2-C1	RAIL	B2	C1		1.85	1.37	
B2	**C1**		**3**		**1.85**	**1.37**	**1.24**
Pelayos	**Aranjuez**					
C2-A	BUS	C2	A		5.10	3.74	
A-A	SUBWAY	A	A		1.50	1.22	
A-C1	RAIL	A	C1		3.40	2.43	
C2	**C1**		**2 C2**		**10.00**	**7.39**	**2.26**
C2-A	BUS	C2	A		5.10	3.74	
A-C1	RAIL	A	C1		3.40	2.43	
C2	**C1**		**2 C2**		**8.50**	**6.17**	**2.26**
Parla	**Pinto**					
B2-B2	RAIL	B2	B2		1.70	1.00	
B2	**B2**		**1**		**1.70**	**1.00**	**0.66**
B2-B2	TRAM	B2	B2		2.00	1.22	
B2	**B2**		**1**		**2.00**	**1.22**	
Pelayos	**Navas del Rey**					
C1-C1	BUS	C1	C1		1.30	0.85	
C1	**C1**		**1**		**1.30**	**0.85**	
C1-B3	BUS	C1	B3		2.00	1.22	
B3-C1	BUS	B3	C1		2.00	1.22	
C1	**B3**		**2**		**4.00**	**2.44**	**1.09**
Yuncos	**Buitragos**					
C1-C1	RAIL	C1	C1		1.70	1.00	
C1	**C1**		**1 C2**		**1.70**	**1.00**	**0.66**
Principe Pio	**Pelayos**					
C2-A	BUS	C2	A		5.10	3.74	
B2	**A**		**C2**		**5.10**	**3.74**	**2.26**
Santa Eugenia	**Aravaca**					
A-A	RAIL	A	A		1.70	1.00	
A	**A**		**A**		**1.70**	**1.00**	**1.24**
Cuzco	**Pinar del Rey**					
A-A	SUBWAY	A	A		1.50	1.22	
A-A	SUBWAY	A	A		1.50	1.22	
A	**A**		**A**		**3.00**	**2.44**	**1.24**
A-A	BUS	A	A		1.50	1.22	
A	**A**		**A**		**1.50**	**1.22**	**1.24**

Figure 4.13: Ticket prices test for different routes

In the previous figure, it is observed that, under the assumption that the people holding a season ticket use PT on a regular basis (weekdays), the cost per trip does not differ much from a single ticket or a 10-trip ticket and it allows the transfer flexibility not allowed by single or 10-trip tickets.

Conclusions

A large majority of PT users have a season ticket. The season ticket is more common among younger people (below 27) and older people (above 64). It is also more common for people who make frequent trips (every weekday). For people between 27 and 64 years old that make around 40-45 trips in a month, in tariff zone A, the 10-ticket and the season ticket have a similar overall monthly cost. This is not the case for younger people and older people, who benefit from a substantial reduction in their season ticket. However, the use of the 10-trips ticket among these groups is much less common. Consequently, assuming a season ticket for all users is reasonable as a first approach.

Figure 4.14 illustrates the different tariff zones. The different season tickets include all the trips that are made in a given zone and any zone inside it. For example, the B1 season ticket includes all trips starting and ending in zones B1 and A. Similarly, the C2 season ticket includes all trips starting and ending in zones C2, C1, B3, B2, B1 and A. To calculate the trip cost, the season ticket price will be divided into the number of trips per month.

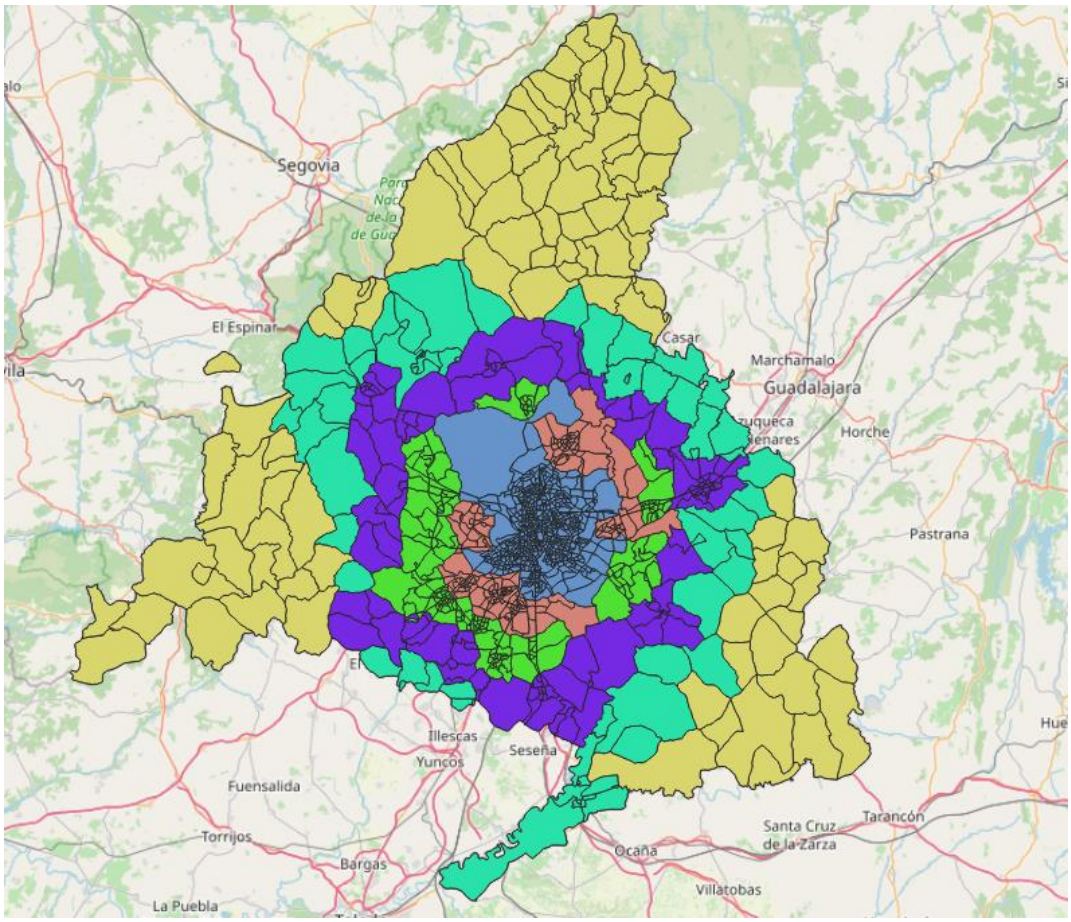


Figure 4.14: Tariff zoning

4.2.3. The logit models

The objective of this component is to calibrate a mode choice function based on the EDM2018 (Madrid region household survey) to segment the demand observed in the mobile phone data by transport mode. In an initial iteration, only three modes will be considered: walk, public transport (PT) and car. Additionally, the demand will be segmented into two demand groups based on the time period (TP) and two multinomial (MNL) logit structures will be tested: 3-MNL, which is a MNL with three options at the same level, and Nested Logit (NL), which contains a nest with the motorised options (public transport and car) and a nest with the active option (walk) (see Figure 4.15).

The logit model models compare the utility function of each mode, which is used to represent individuals' preferences for goods or services. In the transport case, the utility is negative as transport itself presents a disutility to users.

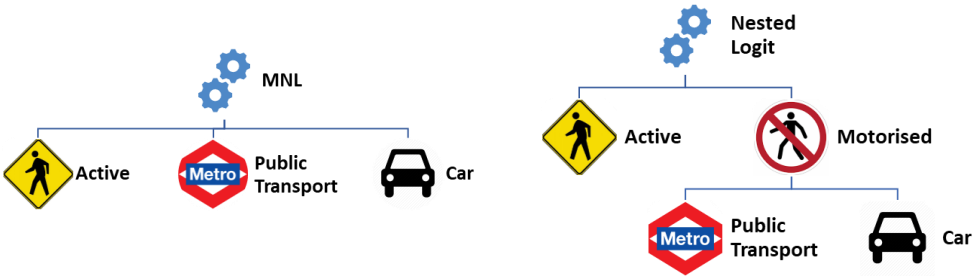


Figure 4.15: Proposed logit model structures

Inputs

The input data to build the modal choice models are the following:

- OD pairs (extracted from the survey)
- Time periods: a set of hours will be grouped by time period (AM: 7:00:00 - 9:59:59, IP: 10:00:00 - 16:29:29, PM: 16:30:00 - 19:29:59, OP: 19:30:00 - 6:59:59). Transport supply will be considered constant along the time period.
- Dates and date groups (weekdays, weekends)
- EDM2018. The survey provides characteristics of the trip and the user making this trip as well as the mode chosen to make this trip. To be able to use this survey, it has been pre-processed. All trips that have an origin or a destination outside Madrid´s region and all trips that have a transport mode which cannot be included in car, public transport and walking have been deleted (for example, taxi and bicycle).

The EDM2018 was collected between 13.02.2018 and 12.06.2018 (only on weekdays). To be able to calculate the routes in OSRM and OTP, a day within the abovementioned dates will be chosen.

Weekends were not included in the EDM2018 survey, hence the model applicable for weekends will be the one calibrated for off-peak on weekdays.

Methodology

Overall workflow

Before starting the calibration of the logit model, the routes need to be calculated as explained in subsection 4.2.1. Subsequently, the component needs to be run with the following inputs:

- OD pairs list resulting from the survey component;
- list of dates and date groups, which ideally should be within the EDM survey period;
- list of hours and hour groups.

As explained above, the routes are not calculated for a given time, but for a day type (weekday, weekend) and an hour type (AM, IP, PM and OP). Once the routes have been calculated, the road and public transport cost and parking information are added to the route characterisation.

This route information will be assigned to the survey data. Initially, the pre-processed survey data will be read. The route information will be then assigned bearing in mind the time period, the origin and the destination.

Subsequently, a cleaning process on the route information has to take place. The reason for this is that some modes may have more than one alternative. There may be several “road”, “walk” or “pt” alternatives making it difficult to compute the costs for each alternative. A simplification is made selecting the route with the lowest travel time within each mode. The generalised cost is computed for that option.

Given this cleaner version of the route extraction, the demand will be segmented into two groups according to their time period:

- peak (including the AM and PM peaks);
- off-peak (including IP and OP)

For each demand group, the following logit functions will be tested:

- Multinomial Logit 3 options (all three options being at the same level)
- Nested Logit (the first level being Active vs Motorised and the motorised branch being split into PT and car)

To calibrate these functions, a maximisation of the maximum log-likelihood is conducted.

These logit functions are calibrated for a given utility function, meaning that different variables will be tested in the utility function. It must be pointed out that only variables that can be derived from mobile phone data will be used to calibrate the logit models. In a first iteration, the utility function will be tested with the disutility of time. Subsequently, the disutility will be tested with time and ticketing/operational costs. Several iterations will take place considering different variables. In each iteration, the overall log-likelihood will be compared and the significance of each variable considered. The addition of variables is evaluated against the improvements in the model accuracy and the explanatory power they add to the model. Once the model is calibrated, the model will be validated with the tests detailed in the validation section.

Biogeme package for Python

To calibrate the different considered models, the Biogeme package (https://biogeme.epfl.ch/) has been used. Biogeme is an open-source Python package designed for the maximum likelihood estimation of parametric models in general, with a special emphasis on discrete choice models. It relies on the package Python Data Analysis Library called Pandas.

Model selection

Mode selection criteria. As explained above, different iterations have been carried out. Each iteration included a different combination of variables and constants. To evaluate each tested model, different parameters have been evaluated:

- the initial and final log-likelihood of each calibrated model;
- the derived general fit index $\rho = 1 - \frac{l(\theta)}{l(0)}$ which explains how much the included variables can explain the observed decisions;
- the sign of each variable and that it is sensible within the model specification (i.e., that the higher the costs, the less attractive an option is);
- the significance of each variable;
- the overall mode share compared to the original survey; and
- the overall accuracy value.

Candidate models. The variables considered for each iteration are detailed in Table 4.36. As for the overall result, these are presented in Table 4.37, Table 4.38, Table 4.39 and Table 4.40, for the MNL peak model, NL peak model, MNL off-peak model and the NL off-peak model respectively.

Table 4.36: Variables used for each model

Model Specification	an ASC for each mode	Time (s) specific to each transport mode	Time (m) specific to each transport mode	Time (ivt) specific to pt	Time (h) specific to each transport mode	dummy distance	dummy young	dummy old	distance walk	road cost	pt cost	parking
Model 0	0	1	0	0	0	0	0	0	0	0	0	0
Model 1	0	1	1	0	0	0	0	0	0	0	0	0
Model 2	0	1	0	1	0	0	0	0	0	0	0	0
Model 3	0	1	0	0	0	1	0	0	0	0	0	0
Model 4	0	1	0	1	1	0	0	0	0	0	0	0
Model 5	0	1	0	1	1	0	0	0	0	0	0	0
Model 6	0	1	0	1	1	0	0	0	0	0	0	0
Model 7	0	1	0	1	1	0	0	1	0	0	0	0
Model 8	0	1	0	1	1	0	0	1	1	0	0	0
Model 9	0	1	0	1	1	0	1	1	1	0	0	0
Model 10	0	1	0	1	1	0	0	1	1	1	0	0
Model 11	0	1	0	1	1	0	0	1	1	1	0	0
Model 12	0	1	0	1	1	0	1	1	1	0	1	0
Model 14	0	1	0	1	1	0	1	1	1	0	1	1
Model 15	0	1	0	1	1	0	1	1	1	0	0	1

Peak models calibration results

Table 4.37: Results for MNL peak model

Model Specification	Initial Log-Likelihood	Final Log-Likelihood	Rho	Overall modal shares and accuracy (baseline modal share: 23.7% walk, 28.8% pt, 47.5% car)
Model 0	-44084	-42378	0.04	0% walk, 0% pt, 100% car, accuracy 47.17%
Model 1	Nan	Nan	Nan	0.03% walk, 0.06% pt, 99.91% car, accuracy 47.13%
Model 2	-595388	-296065	0.50	20.5% walk, 28.85% pt, 50.65% car, accuracy 54.85%
Model 3	-37023	-36671	0.01	31.26% walk, 1.47% pt, 67.27% car, accuracy 58.52%
Model 4	-44084	-36091	0.18	35.83% walk, 9.32% pt, 54.85% car, accuracy 56.8%
Model 5	-572804	-230212	0.60	24.43% walk, 28.71% pt, 46.86% car, accuracy 56.95%
Model 6	-44084	-35499	0.19	35.33% walk, 7.78% pt, 56.89% car, accuracy 57.32%
Model 7	-572804	-228634	0.60	24.28% walk, 28.73% pt, 47.0% car, accuracy 57.05%
Model 8	-572804	-228288	0.60	24.29% walk, 28.72% pt, 46.99% car, accuracy 57.2%
Model 9	-572804	-219053	0.62	24.17% walk, 28.77% pt, 47.05% car, accuracy 56.7%
Model 10	-572804	-228258	0.60	24.29% walk, 28.72% pt, 46.99% car, accuracy 57.2%
Model 11	-572804	-228256	0.60	24.29% walk, 28.72% pt, 46.99% car, accuracy 57.2%
Model 12	-572804	-225956	0.61	24.35% walk, 28.71% pt, 46.94% car, accuracy 58.02%
Model 13	-572804	-217236	0.62	24.35% walk, 28.65% pt, 47.0% car, accuracy 60.23%
Model 14	-572804	-217632	0.62	24.39% walk, 28.66% pt, 46.96% car, accuracy 60.0%
Model 15	-572804	-216247	0.62	24.45% walk, 28.68% pt, 46.88% car, accuracy 61.09%

Table 4.38: Results for NL peak model

Model Specification	Initial Log-Likelihood	Final Log-Likelihood	Rho	Overall modal shares and accuracy (baseline modal share: 23.7% walk, 28.8% pt, 47.5% car)
Model 0	-44211	-42378	0.02	0% walk, 0% pt, 100% car, accuracy 47.17%
Model 1	Nan	Nan	Nan	0.01% walk, 28.23% pt, 71.76% car
Model 2	-617460	-296065	0.52	20.5% walk, 28.85% pt, 50.65% car, accuracy 54.85%
Model 3	-37045	-36671	0.01	31.26% walk, 1.47% pt, 67.27% car, accuracy 58.52%
Model 4	-44211	-36091	0.18	35.83% walk, 9.32% pt, 54.85% car, accuracy 56.8%
Model 5	-591738	-230212	0.61	24.43% walk, 28.71% pt, 46.86% car, accuracy 56.95%
Model 6	-44211	-35499	0.20	35.33% walk, 7.78% pt, 56.89% car, accuracy 57.32%
Model 7	-591738	-228634	0.61	24.28% walk, 28.73% pt, 47.0% car, accuracy 57.05%
Model 8	-591738	-228288	0.61	24.28% walk, 28.73% pt, 46.99% car, accuracy 57.19%
Model 9	-591738	-219053	0.63	24.17% walk, 28.77% pt, 47.05% car, accuracy 56.7%
Model 10	-591738	-228258	0.61	24.29% walk, 28.72% pt, 46.99% car, accuracy 57.2%
Model 11	-591738	-228256	0.61	24.29% walk, 28.72% pt, 46.99% car, accuracy 57.2%
Model 12	-591738	-225956	0.62	24.35% walk, 28.71% pt, 46.94% car, accuracy 58.02%
Model 13	-591738	-217236	0.63	24.35% walk, 28.65% pt, 47.0% car, accuracy 60.23%
Model 14	-591738	-217632	0.63	24.39% walk, 28.66% pt, 46.96% car, accuracy 60.0%
Model 15	-591738	-216247	0.63	24.45% walk, 28.68% pt, 46.88% car, accuracy 61.09%

Off-peak models calibration results

Table 4.39: Results for MNL off-peak model

Model Specification	Initial Log-Likelihood	Final Log-Likelihood	Rho	Overall modal shares and accuracy (baseline modal share: 26.6% walk, 31.1% pt, 43.2% car)
Model 0	-52069	-51030	0.02	100% car, accuracy 42.73%
Model 1	Nan	Nan	0.04	0.06% walk, 0.01% pt, 99.84% car, accuracy 42.89%
Model 2	-764395	-371829	0.51	20.6% walk, 30.94% pt, 48.46% car, accuracy 55.08%
Model 3	-44893	-43821	0.02	40.28% walk, 8.97% pt, 50.75% car, accuracy 56.58%
Model 4	-52069	-43124	0.17	39.5% walk, 17.88% pt, 42.62% car, accuracy 57.66%
Model 5	-724633	-285542	0.61	26.23% walk, 31.12% pt, 42.65% car, accuracy 57.72%
Model 6	-52069	-42163	0.19	39.25% walk, 21.54% pt, 39.2% car, accuracy 57.84%
Model 7	-724633	-282811	0.61	26.15% walk, 31.25% pt, 42.61% car, accuracy 58.06%
Model 8	-724633	-282114	0.61	26.16% walk, 31.19% pt, 42.65% car, accuracy 58.38%
Model 9	-724633	-270642	0.63	26.1% walk, 31.2% pt, 42.7% car, accuracy 58.07%
Model 10	-724633	-282077	0.61	26.16% walk, 31.18% pt, 42.65% car, accuracy 58.38%
Model 11	-724633	-282074	0.61	26.16% walk, 31.2% pt, 42.65% car, accuracy 58.38%
Model 12	-724633	-279549	0.61	26.26% walk, 31.13% pt, 42.65% car, accuracy 59.15%
Model 13	-724633	-270606	0.63	26.19% walk, 31.3% pt, 42.5% car, accuracy 60.91%
Model 14	-724633	-271104	0.63	26.25% walk, 31.32% pt, 42.43% car, accuracy 60.73%
Model 15	-724633	-270711	0.63	26.24% walk, 31.26% pt, 42.5% car, accuracy 61.2%

Table 4.40: Results for NL off-peak model

Model Specification	Initial Log-Likelihood	Final Log-Likelihood	Rho	Overall modal shares and accuracy (baseline modal share: 26.6% walk, 31.1% pt, 43.2% car)
Model 0	-52190	-51030	0.02	100% car, accuracy 42.73
Model 1	Nan	Nan	Nan	0.03% walk, 27.6% pt, 72.36% car, accuracy 41.8%
Model 2	-793155	-371829	0.53	20.6% walk, 30.94% pt, 48.46% car, accuracy 55.08%
Model 3	-44950	-43821	0.03	40.28% walk, 8.96% pt, 50.75% car, accuracy 56.57%
Model 4	-52190	-43124	0.17	39.5% walk, 17.88% pt, 42.62% car, accuracy 57.66%
Model 5	-748914	-285542	0.62	26.23% walk, 31.12% pt, 42.65% car, accuracy 57.72%
Model 6	-52190	-42163	0.19	39.25% walk, 21.54% pt, 39.2% car, accuracy 57.84%
Model 7	-748914	-282811	0.62	26.15% walk, 31.25% pt, 42.61% car, accuracy 58.06%
Model 8	-748914	-282114	0.62	26.16% walk, 31.19% pt, 42.65% car, accuracy 58.38%
Model 9	-748914	-270642	0.64	26.1% walk, 31.2% pt, 42.7% car, accuracy 58.07%
Model 10	-748914	-282077	0.62	26.16% walk, 31.18% pt, 42.65% car, accuracy 58.38%
Model 11	-748914	-282074	0.62	26.16% walk, 31.18% pt, 42.65% car, accuracy 58.38%
Model 12	-748914	-279549	0.63	26.26% walk, 31.13% pt, 42.65% car, accuracy 59.15%
Model 14	-748914	-270606	0.64	26.19% walk, 31.3% pt, 42.5% car, accuracy 60.91%
Model 15	-748914	-271104	0.64	26.25% walk, 31.32% pt, 42.43% car, accuracy 60.73%

4.3. Validation plan

4.3.1. Overview

The calibrated model has been subsequently validated to ensure that the model produces sensible results and that the resulting model is not overfitted.

4.3.2. Objectives

Since it is currently very difficult to identify the transport mode from mobile phone data in urban areas, we have calibrated a logit model to estimate the modal choices of the trips detected from the mobile network data. In subsequent project stages, this information will be used as an input to the long-distance travel simulation models.

4.3.3. Validation approach

The validation approach is summarised below:

- Divide the used data into training and test data
- Divide the training data into different subgroups to cross-validate the training data:
    - Calibrate the logit model maximising the log-likelihood of the model excluding one of the subgroups.
    - Apply the resulting model to the previously excluded subgroup.
    - Compare the resulting log-likelihoods of all the resulting combinations (similar values are to be expected).
- Validate the resulting model with the test data:
    - The resulting mode shares should be similar to the observed data.
    - The resulting mode shares should be similar to the calibrated values.
- Validate the model relevance by comparing the first preference recovery against the expected recovery

4.3.4. Data and software inputs

To calibrate and validate the proposed mode identification approach, household surveys as well as transport supply data obtained from OSRM and OTP have been used. To clean the household data the following steps have been taken:

- the surveys with an origin or destination outside the study area have been filtered out;
- the surveys with the same origin and destination have been filtered out; and
- the surveys with a mode not considered by the model have been filtered out

The parameters used for the validation experiments are the following:

- training test split: 50/50;
- cross-validation split: 4 folds

4.4. Results

All the survey cases that are in the pre-processed sample are randomly assigned a value and according to this value are assigned to calibration or validation. For the EDM 2018 calibration the split used for calibration and validation has been 50/50. The table below compares the modal shares achieved with the model with the ones derived from the survey both for calibration and validation samples. Neither intrazonal trips nor trips in other modes rather than walking, public transport and car are considered.

Table 4.41: Aggregated mode shares (%)

	Survey			MNL			NL		
	Car	Walk	PT	Car	Walk	PT	Car	Walk	PT
Peak Calibration	47.50	23.70	28.80	46.96	24.39	28.66	46.96	24.39	28.66
Peak Validation	47.50	23.70	28.80	46.92	24.79	28.29	46.92	24.79	28.29
Off-peak Calibration	43.20	25.60	31.10	42.43	26.25	31.32	42.43	26.25	31.32
Off-peak Validation	43.20	25.60	31.10	42.23	26.43	31.34	42.23	26.43	31.34

When it comes to the aggregated mode shares, it is observable for both peak and off-peak that the estimated mode shares do not differ significantly between calibration and validation. More importantly, there is no significant difference between the calculated mode shares and the ones from the original survey.

4.4.1. MN model vs NL model

When analysing the model parameters, it is observed that the lambda that differentiates the NL from the MNL has an estimated value of 1. Thus, both estimated models are similar. Consequently, in any further analysis, there will be no reference to the NL model.

4.4.2. Model parameters

Taking a closer look at the parameters, it can be observed that time in a car is more penalised than walking, whereas time in PT is penalised less than walking.

Table 4.42: Variables results for the peak MNL model

Variable Name	Value	Std err	t-test	p-value	Rob. Std err	Rob. t-test	Rob. p-value
ASC car (car mode constant)	-13.2	0.0661	-200	0	0.258	-51.3	0
ASC pt (public transport constant)	29.9	0.302	99.2	0	1.17	25.6	0
BETA cost pt (parameter penalising pt cost)	-23.6	0.202	-117	0	0.766	-30.8	0
BETA dummy old (parameter penalising being +65)	1.94	0.0838	23.2	0	0.228	8.52	0
BETA dummy short (parameter penalising being less than 5km)	29.6	0.149	199	0	0.445	66.5	0
BETA dummy young (parameter penalising being under 27)	6.62	0.0961	68.8	0	0.338	19.5	0
BETA time car (parameter penalising car time)	-2	0.00999	-200	0	0.0504	-39.6	0
BETA time pt (parameter penalising pt time)	-0.528	0.00488	-108	0	0.0229	-23.1	0

Table 4.43: Variables results for the off-peak MNL model

Variable Name	Value	Std err	t-test	p-value	Rob. Std err	Rob. t-test	Rob. p-value
ASC car (car mode constant)	-12.7	0.0687	-185	0	0.246	-51.8	0
ASC pt (public transport constant)	31.3	0.31	101	0	1.1	28.4	0
BETA cost pt (parameter penalising pt cost)	-24.1	0.207	-116	0	0.721	-33.5	0
BETA dummy old (parameter penalising being +65)	1.82	0.107	17	0	0.291	6.25	4.1e-10
BETA dummy short (parameter penalising being less than 5km)	29.8	0.159	188	0	0.462	64.6	0
BETA dummy young (parameter penalising being under 27)	5.18	0.0897	57.8	0	0.287	18	0
BETA time car (parameter penalising car time)	-1.99	0.0106	-189	0	0.0483	-41.3	0
BETA time pt (parameter penalising pt time)	-0.566	0.00544	-104	0	0.0239	-23.7	0

4.4.3. Model sensitivity

The probability distributions of the estimated and actual choices have been analysed in order to check the model sensitivity. First, the probability of the estimated mode is evaluated. This analysis aims at checking the probability of the best choice according to the model. Figure 4.16 and Figure 4.17 show that over 99% of the cases have a probability over 50%. This means that the second and third options had a very low probability compared to the best option.

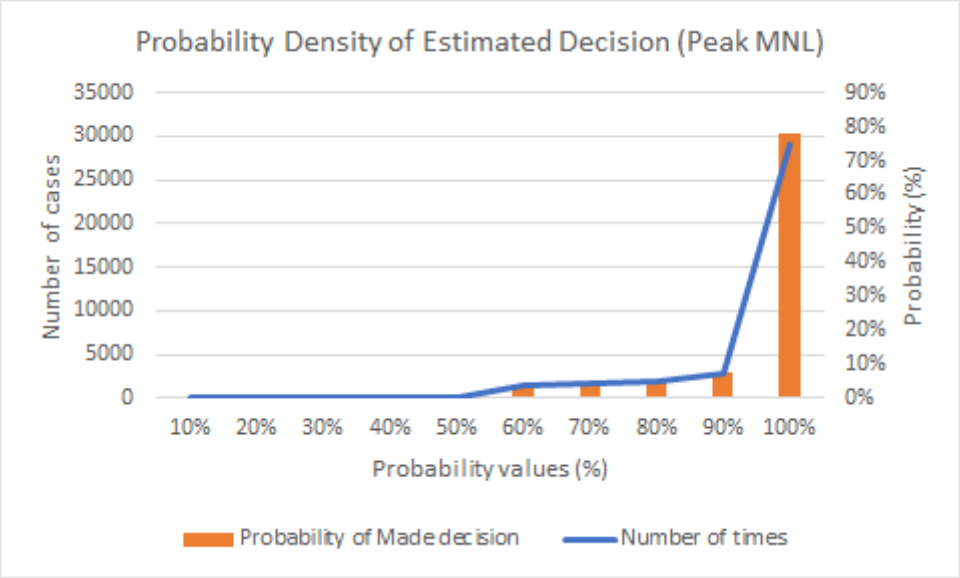


Figure 4.16: Probability density of estimated decisions in peak MNL model (Calibration)

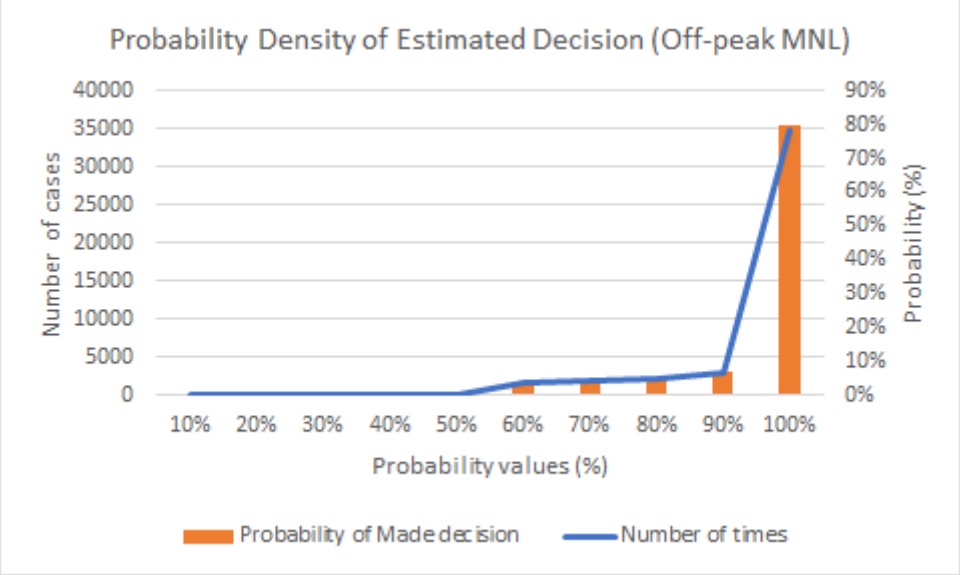


Figure 4.17: Probability density of estimated decisions in off-peak MNL model (Calibration)

Second, the probability distribution of the observed decision is evaluated. Figure 4.18 and Figure 4.19 show the results. A bimodal distribution can be observed. Most decisions correctly estimated by the model are associated with very high probabilities, while most decisions incorrectly estimated by the model are associated with very low probabilities.

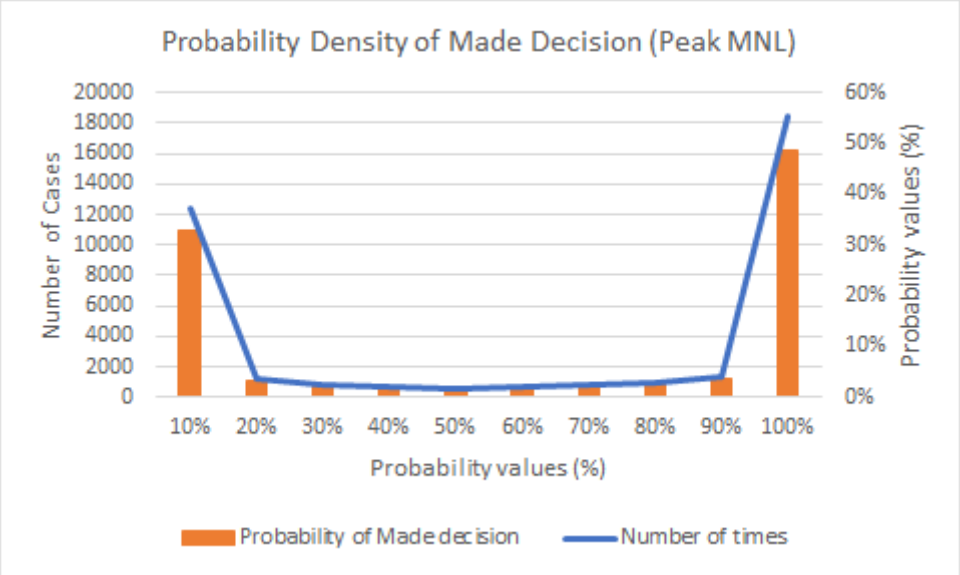


Figure 4.18: Probability density of actual decision in peak MNL model (Calibration)

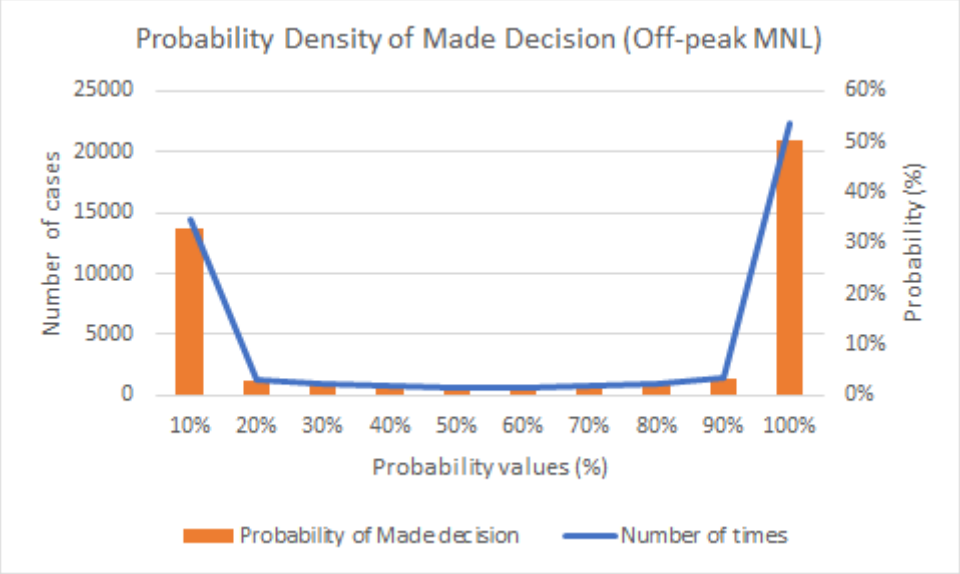


Figure 4.19: Probability density of actual decision in off-peak MNL model (Calibration)

This is most likely due to the estimated parameters of the model. The values of the constants and dummy variables take very high values compared to the ones that are multiplied by time (Table 4.42 and Table 4.43). In other words, the calibrated model is more dependent on certain characteristics of the trip or the traveller than on the travel time.

4.4.4. Model relevance

To ensure that the model is relevant, the following tests have been carried out:

- **First Preference Recovery** (FPR), which is the number of individuals correctly predicted.
- **Chance Recovery** (CR), which is the number of individuals that would be assigned to each mode based on random probabilities (1 divided by the number of alternatives), and subsequently adding all the cases.
- **Expected Recovery** (ER), which is based on the maximum probability for each case, and subsequently adding all the probabilities.

From these estimates, two hypotheses are to be checked:

- FPR and ER should be similar. For this, FPR, ER and its variance are calculated. Subsequently, using the normal distribution values, a confidence interval for ER is defined and compared against the FPR value.
- FPR should be significantly greater than CR.

Table 4.44: Relevance test for the MNL peak and off-peak models

	Peak model	Off-peak model
Portion of correctly predicted	0.60	0.61
FPR	24,077 (out of 40,127)	28,781 (out of 47,395)
ER	(37,613 – 37,749)	(44,521 – 44,666)
CR	(8,727-9,107)	(10,326-10,738)

When comparing the FPR and the ER, it can be said that FPR and ER are not similar for a confidence interval of 95%. This is due to the following facts. Firstly, the average ER is significantly higher than the FPR as the average probability for the “most likely option” is rather high. Secondly, the interval based on the ER variance is rather short given the small variance in the “most likely option” probabilities. On the other hand, when comparing the FPR and the ER, it can be said that the FPR value is clearly higher than the CR.

4.4.5. Application to modal shares across OD pairs between tariff zones

On top of the overall mode share comparisons, mode shares have been estimated for different demand segmentations. The tested demand segmentations are based on the tariff zone where the trip starts and ends. It is believed and subsequently confirmed in the data that mode shares will vary significantly depending on the OD pair. As it can be observed in Table 4.45 and Table 4.46, trips starting and ending in the city centre (crown CA1 corresponds to the intersection between tariff zone A and whatever it is inside the M30 ring road) have a walking mode share of almost 50% and a PT mode share of 38% and 39% for peak and off-peak respectively. Similar results have been derived from the calibrated logit model, with an estimated walking mode share of almost 50% and a PT mode share of around 30%, slightly under the observed data.

Regarding trips starting in crown C and ending in crown B, the respondents are clearly inclined to take the car, with a mode share of over 80%. The estimated results overestimate the car share and underestimate the PT mode share. This might be due to worse PT data in areas far from the city centre, as the underground and certain rail services are well represented but some intercity buses might not be as up to date as the other services.

Table 4.45: Peak mode share per aggregated tariff zone (survey and calibration and validation results)

Peak	Survey			Calibration			Validation		
Tariff Crowns	Car	Walk	PT	Car	Walk	PT	Car	Walk	PT
CA1-CA1	15%	47%	38%	23%	49%	28%	19%	49%	32%
CA1-CA2	36%	9%	55%	41%	7%	52%	33%	6%	62%
CA2-CA1	29%	7%	63%	40%	6%	54%	32%	7%	61%
CA2-CA2	39%	36%	26%	35%	39%	25%	30%	42%	28%
CA1-CB	51%	0%	49%	45%	0%	55%	39%	0%	61%
CB-CA1	43%	0%	57%	38%	0%	62%	37%	0%	63%
CA2-CB	72%	1%	27%	66%	1%	33%	59%	1%	40%
CB-CA2	69%	0%	31%	64%	0%	36%	60%	0%	40%
CB-CB	57%	31%	12%	54%	29%	17%	50%	32%	17%
CA1-CC	54%	0%	46%	63%	0%	37%	69%	0%	31%
CA2-CC	73%	0%	27%	75%	0%	25%	75%	0%	25%
CB-CC	86%	1%	13%	94%	2%	4%	96%	0%	4%
CC-CC	72%	14%	13%	90%	10%	0%	93%	7%	0%
CC-CA1	53%	0%	47%	61%	0%	39%	59%	0%	41%
CC-CA2	65%	0%	35%	76%	0%	24%	82%	0%	18%
CC-CB	81%	0%	18%	94%	2%	4%	95%	1%	4%

Table 4.46: Off-peak mode share per aggregated tariff zone (survey and calibration and validation results)

Off-peak	Survey			Calibration			Validation		
Tariff Crowns	Car	Walk	PT	Car	Walk	PT	Car	Walk	PT
CA1-CA1	11%	49%	39%	18%	50%	32%	21%	48%	30%
CA1-CA2	26%	8%	66%	34%	7%	59%	37%	7%	56%
CA2-CA1	31%	8%	60%	32%	6%	62%	39%	6%	55%
CA2-CA2	33%	38%	29%	30%	42%	28%	36%	39%	25%
CA1-CB	42%	0%	57%	39%	0%	61%	46%	0%	54%
CB-CA1	48%	0%	52%	35%	0%	65%	38%	0%	62%
CA2-CB	66%	0%	33%	63%	0%	37%	67%	1%	32%
CB-CA2	69%	1%	30%	59%	0%	41%	62%	1%	38%
CB-CB	53%	34%	14%	50%	32%	18%	53%	30%	17%
CA1-CC	47%	0%	53%	63%	0%	37%	69%	0%	31%
CA2-CC	64%	0%	36%	77%	0%	23%	77%	0%	23%
CB-CC	80%	0%	19%	93%	1%	6%	95%	1%	4%
CC-CC	68%	17%	14%	89%	11%	0%	90%	10%	0%
CC-CA1	49%	0%	51%	65%	0%	35%	66%	0%	34%
CC-CA2	69%	0%	31%	71%	0%	29%	76%	0%	24%
CC-CB	83%	1%	16%	95%	1%	4%	94%	2%	3%

4.5. Application to modal share in airport access

4.5.1. Initial results

The mode share to access the Madrid-Barajas airport has been tested. The analysis covers access by private car, PT and walking. It must be pointed out that for this case in particular (access and egress to/from the Airport), the taxi share is significant. The limited modal share achieved by this mode in the whole city makes it not viable to calibrate the model including taxi modes. Specific access mode surveys (e.g., EMMA in Barajas case) could help to overcome this limitation.

The mode share according to the survey is the following:

Table 4.47: Access and egress from and to the airport according to the EDM2018 survey

Access or egress	Car Share (%)	PT Share (%)	Walk Share (%)	Number of cases
Access to the airport	82%	17%	1%	472
Egress from the airport	81%	17%	2%	482

The same probabilities have been estimated with the peak and off-peak model. The model fails to capture PT, as well as certain characteristics that are particularly relevant for airport access (e.g., the agents accessing and egressing the airport are sometimes carrying bag, which makes walking less appealing).

Table 4.48: Access and egress from and to the airport for according to the peak MNL calibration mode

Access or egress	Car Share (%)	PT Share (%)	Walk Share (%)	Number of cases
Access to the airport	88%	0%	12%	78
Egress from the airport	93%	0%	7%	61

Table 4.49: Access and egress from and to the airport for according to the off-peak MNL calibration mode

Access or egress	Car Share (%)	PT Share (%)	Walk Share (%)	Number of cases
Access to the airport	89%	0%	11%	155
Egress from the airport	88%	0%	12%	172

4.5.2. Model sensitivity to centroid location

The singular results for PT were further analysed to check if any inconsistency in the model input may be causing a 0% modal share for this option. Two cases from the survey that reported choosing PT were analysed: (i) a trip from the airport to Latina district, and (ii) a trip from Avenida de América to the airport.

Case 1: Airport to Latina district

Figure 4.20 and Figure 4.21 show the origin and destination zones of the trip and the route recommendations from Google Maps.

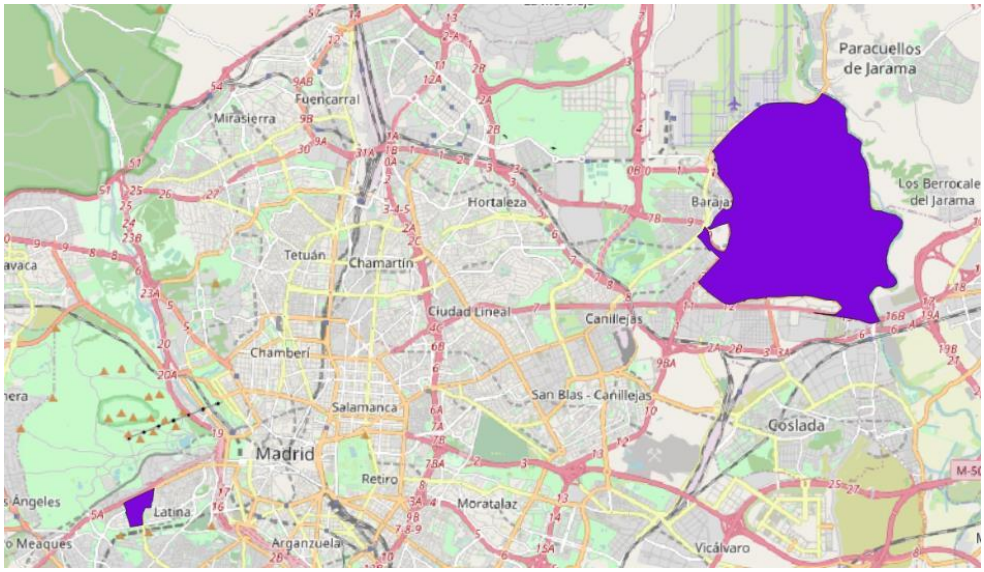


Figure 4.20: From Airport to Latina

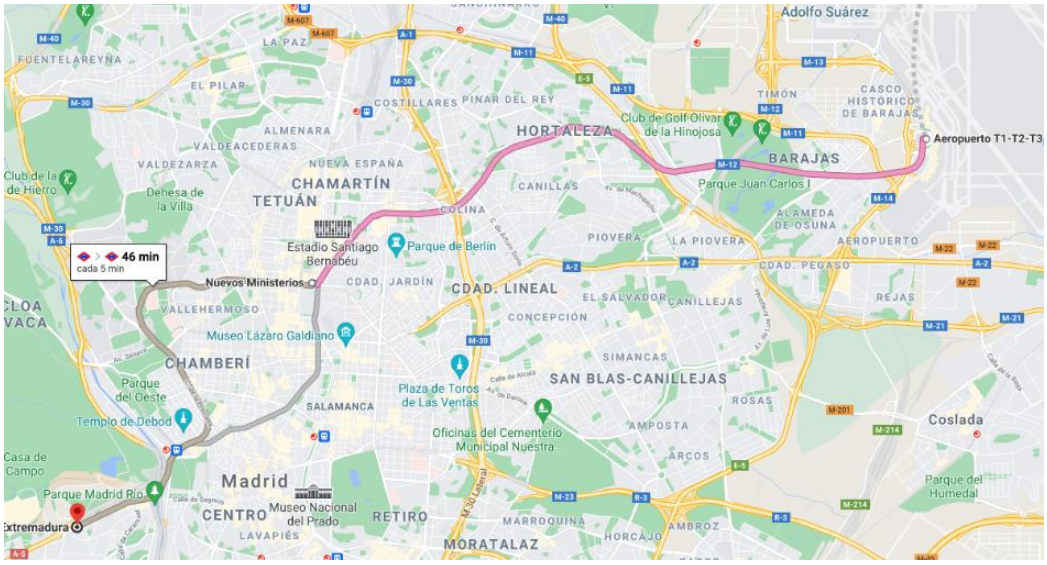


Figure 4.21: From Airport to Latina Google maps query (by public transport)

The results from the routes extraction are summarised below:

- Road time: 2,252 seconds
- PT time: 6,049 seconds (walk) + 1,920 seconds ivt)
- Walk time: 13,134 seconds

In a closer analysis the first leg of the public transport characterisation is illustrated below:

- Duration: 5,086 seconds
- Distance: 6,515 metres
- Stop: 3:par_5_71
- Mode: walk

It is observed that the traveller has to walk over 6 km to reach the best public transport alternative. This means that the centroid location is affecting this value and therefore impacting the predictive ability of the model.

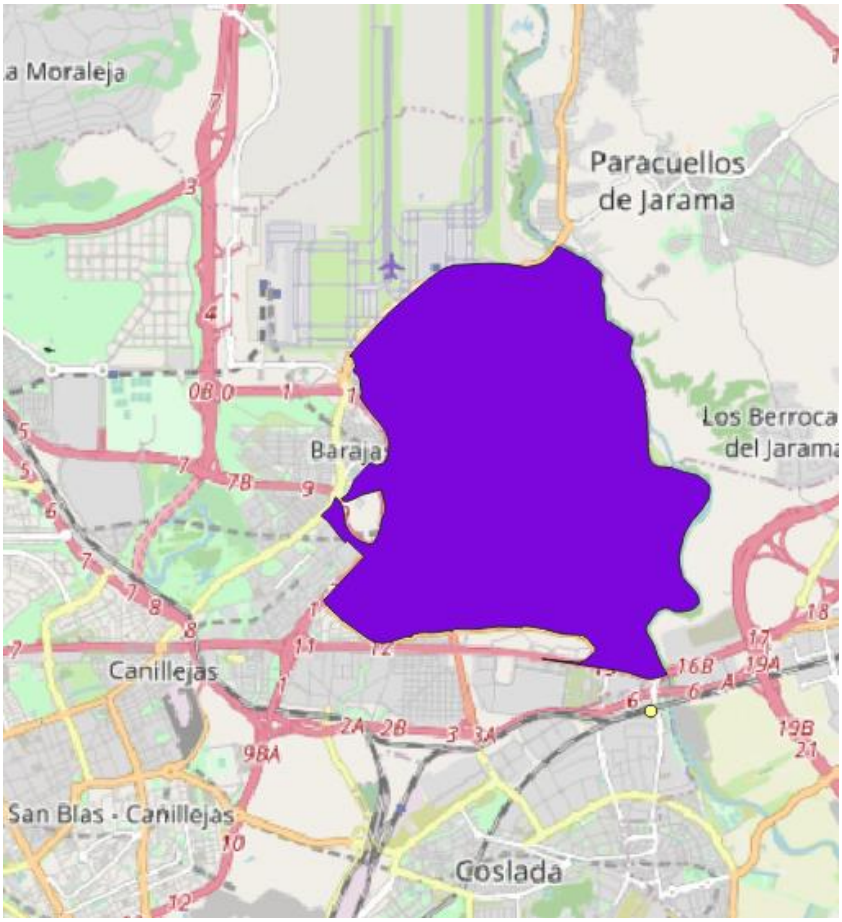


Figure 4.22: From Airport to the recommended closest station

Case 2: Airport to Avenida America

Figure 4.23 and Figure 4.24 show the origin and destination zones of the trip and the route recommendations from Google Maps.

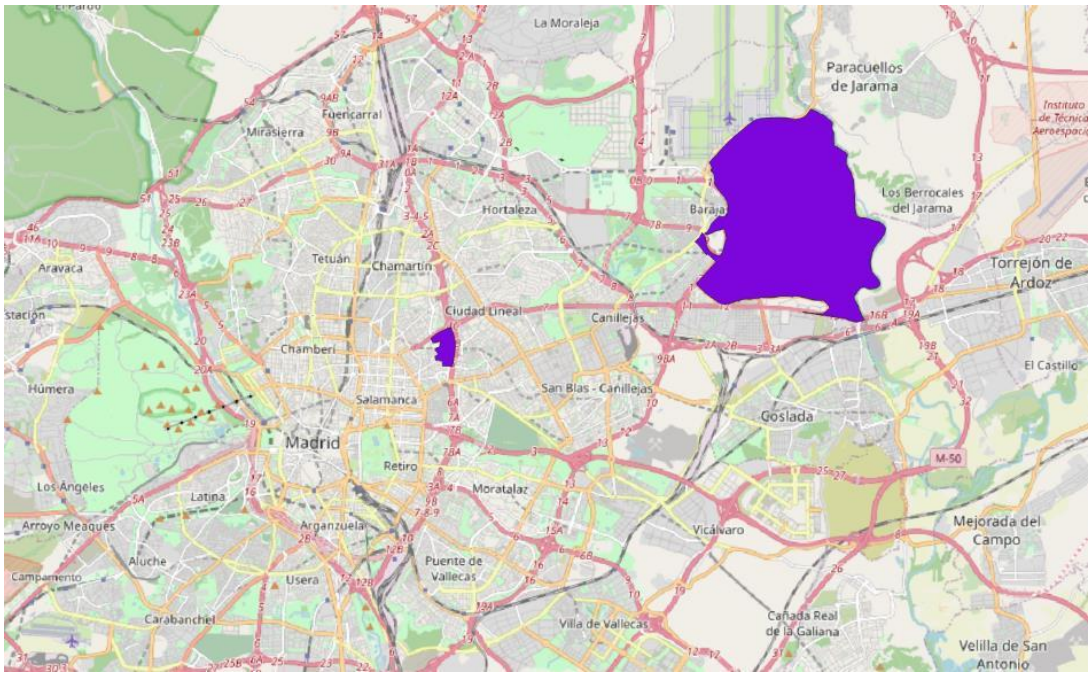


Figure 4.23: From Airport to Avenida America

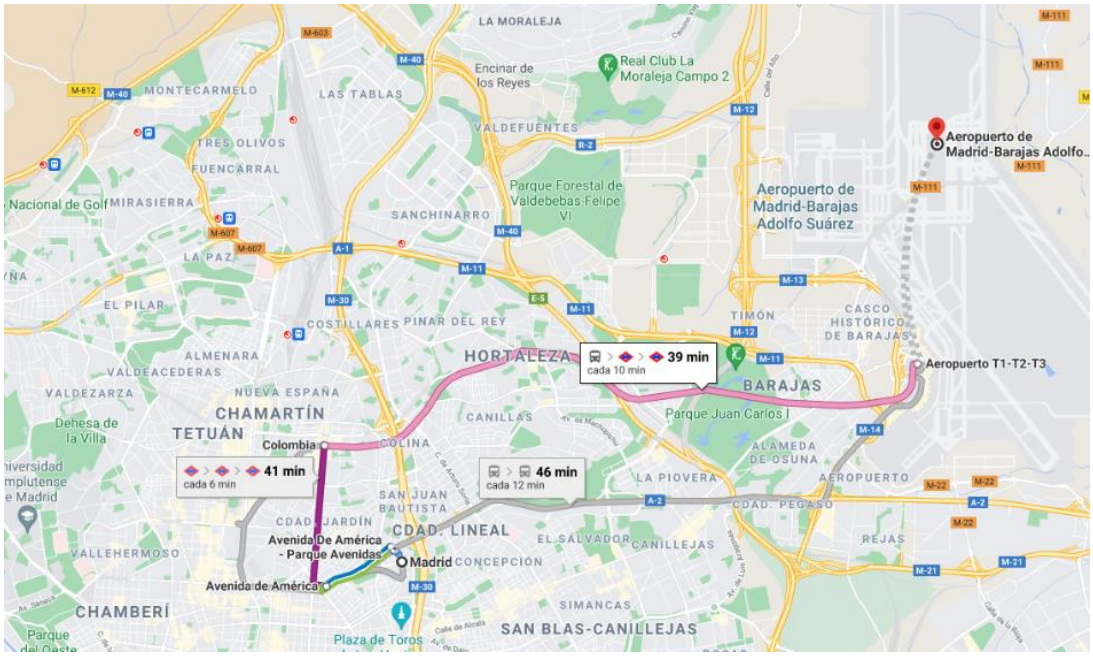


Figure 4.24: From Airport to Avenida America to Airport Google maps query (by public transport)

The results from the routes extraction are summarised below:

- Road time: 1,445 seconds
- PT time: 5,119 seconds (walk) + 1,277 seconds ivt)
- Walk time: 6,807 seconds

In a closer analysis the last leg of the public transport characterisation is illustrated below:

- Duration: 4,867 seconds
- Distance: 6,317 metres
- Stop: 5:par_8_07208
- Mode: walk

It is observed that the user has to walk almost 5 km to reach the best public transport alternative, leading to a similar conclusion than the Airport-Latina case.

4.5.3. Final results

These big walking distances to reach the "best" public transport alternative are due to the fact that the zone centroid is used to calculate the trip characteristics. In this particular case, the zones are quite big and the centroid is located far from the main public transport stops and stations. Therefore, in a second iteration, the points representing the zone have been moved to a more realistic location and the routes have been recalculated.

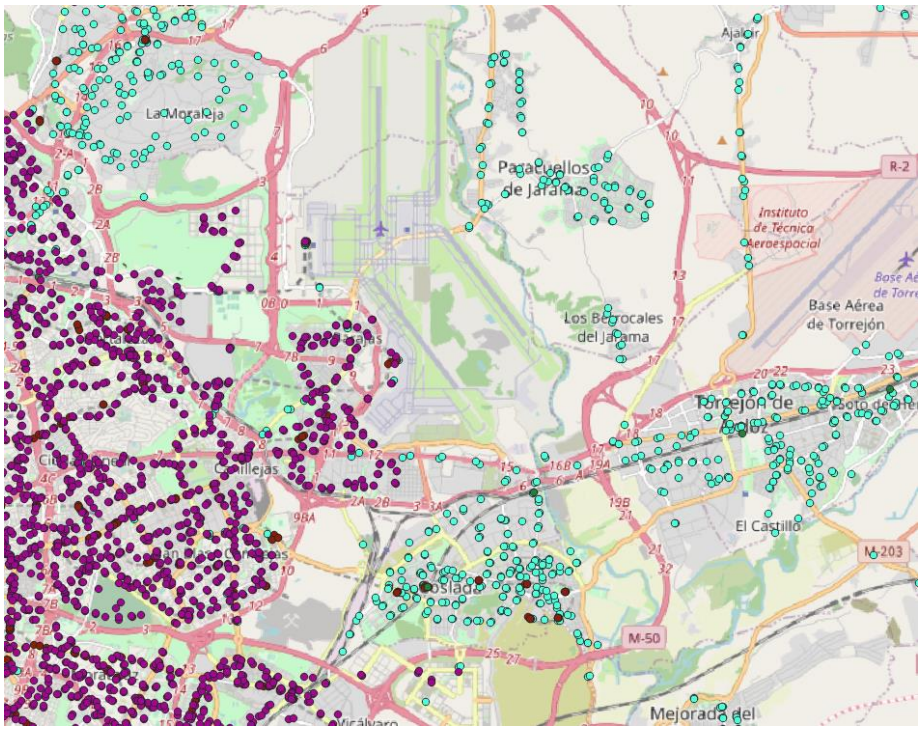


Figure 4.25: Available stops and stations within the Airport surroundings

For this second iteration, the results were as follows:

Table 4.50: Access and egress from and to the airport for according to the peak MNL calibration mode (with improved centroid locations)

Access or egress	Car Share (%)	PT Share (%)	Walk Share (%)	Number of cases
Access to the airport	59%	40%	1%	78
Egress from the airport	52%	46%	2%	61

Table 4.51: Access and egress from and to the airport for according to the off-peak MNL calibration mode (with improved centroid locations)

Access or egress	Car Share (%)	PT Share (%)	Walk Share (%)	Number of cases
Access to the airport	64%	33%	3%	153
Egress from the airport	56%	41%	2%	172

4.6. Conclusion

The model has good results when it comes to aggregated mode shares, both at an individual and tariff zone level. However, when focusing on airport access, the final model overestimates the modal share of PT. To improve these results, in subsequent project stages the model shall be updated to include specific variables dealing with the trip purpose to distinguish between a regular trip and the access to the airport, which has clearly differentiated characteristics.

5. Long-distance travel purpose

5.1. Problem statement

The aim of this section is to report the results of the development of a machine learning model to predict whether a flight trip has a business purpose or another purpose.

5.2. Methodology

5.2.1. Dataset

The dataset considered for this experiment is the EMMA survey performed to travellers in the Madrid Barajas airport during two months of 2018 (June and November). This survey asks travellers at the airport about their trips, whether they are going or returning, the purpose of the trip as well as many other trip-related (flight ticket type, accommodation at destination, etc.) and sociodemographic details (age, gender, studies, etc.).

This dataset contains a total 18,705 responses from travellers performing trips during summer (June 2018) and winter (November 2018) periods. The share of responses between both seasons is of 51% and 49% respectively.

Data preparation and filtering criteria

In general terms, each feature in the dataset has been cleaned and pre-processed according to their types. The basic data cleaning and filtering methods that have been applied are the following:

- Missing data has been managed differently according to features. Missing data in relevant features has been solved by removing the cases involved.
- For simplicity, connection flights have been filtered out of the training data.
- Categorical variables with more than two categories have been converted to one-hot encoding, which consists in the generation of a new feature per category set to 1 when the category is observed and zero otherwise. When possible, the number of categories has been reduced by grouping them together.
- Categorical variables with only two categories have been encoded using a single binary variable.
- Dates have been converted into the corresponding day of the week (0: Monday to 6: Sunday).

From this data cleaning process, 44.4% of the data has been removed, 24.6% corresponding to value cleaning and 19% corresponding to connection filtering.

Feature list

After data cleaning, model input features have been derived from the available features. Table 5.52 below contains a relation of the model features that have been developed, their type, description and the calculation methodologies as well as the EMMA variable (or variables) they have been derived from.

Table 5.52: List of derived features with their description, calculation and corresponding EMMA variables.

Feature name	EMMA correspondence	Description	Calculations	Type	Available in mobile phone data
taus	taus	The length of the stay in days	Basic cleaning	Integer	Yes
working	cdslab	1.0 if the respondent works 0.0 otherwise	From the cdslab variable identify those respondents that declared being working.	Binary	No
acompt2	acompt2	The number of people that go to the airport with the traveller	Basic cleaning	Integer	No
npers	npers	The number of people that travel with the respondent	Basic cleaning	Integer	No
gender	cdsexo	The gender of the respondent	1 for female and 0 for male	Binary	Yes
arrival_weekday	fllegada, cdidavue, taus	The weekday of the arrival date (0 monday 6 sunday). Arrival date is the reported fllegada if cdidavue is “ida” and the updated fllegada with taus if cdidavue is “vuelta”.	Compute arrival and return dates for all trips depending of cdidavue and taus variables and update arrival_weekday with the observed values for the “ida” case and the computed values for the “vuelta” case.	Integer	Yes
departure_weekday	fllegada, cdidavue, taus	The weekday of the departure date (0 monday 6 sunday). Departure date is the reported fllegada if cdidavue is “vuelta” and the updated fllegada with taus if cdidavue is “ida”.	Compute arrival and return dates for all trips depending of cdidavue and taus variables and update arrival_weekday with the observed values for the “vuelta” case and the computed values for the “ida” case.	Integer	Yes
arrival_hour	hllega	The reported time of arrival.	The hour is truncated by ignoring the minutes. Invalid hour values are removed	Integer	Yes
month categories	mes	A categorical variable for the month of the trip	There is a month variable that gets categorised using one-hot encoding	Categorical	Yes
days_rel_festivity	fllegada, cdidavue	The days to/from festivity depending on whether the trip is “ida” or “vuelta”	Compute days to festivity and days from festivity and select the former when cdidavue is “ida” and the latter for “vuelta.	Integer	Yes

Feature name	EMMA correspondence	Description	Calculations	Type	Available in mobile phone data
approximate_age	cdedad	The approximated age of each survey respondent	Compute the average age of the boundaries of the group the user is assigned to (20 to 25 is 22.5)	Float	No
age_E0	cdedad	User belongs to the E0 age group (15 to 19 years old)	Convert the group 15 to 19 defined in cedad into E0	Binary	Yes
age_E1	cdedad	User belongs to the E1 age group (20 to 49 years old)	Merge groups 20 to 24, 25 to 29, 30 to 39 and 40 to 49 into E1	Binary	Yes
age_E2	cdedad	User belongs to the E2 group (50 to 64 years old)	Merge groups 50 to 59 and 60 to 64 into E2	Binary	Yes
age_E3	cdedad	User belongs to the E3 group (65 years old or older)	Merge all remaining groups into E3	Binary	Yes
residence_region	cdpaisre	The region where the respondent lives	Convert all countries into Spain, Europe, Latinamerica and rest of the world	Categorical	Yes
nationality_region	cdpaisna	The region where the respondent comes from	Convert all countries into Spain, Europe, Latinamerica and rest of the world	Categorical	Yes
mode categories	modos	The mode each respondent arrives to the airport	A categorical dummy variable for each possible mode defined	Categorical	No
accommodation categories	cdalojin	The declared accommodation declared by each survey respondent.	A categorical one-hot encoded variable for each possible accommodation.	Categorical	No
education categories	estudios	The declared education each respondent has achieved.	A categorical one-hot encoded variable for each possible education level.	Categorical	No
ticket categories	billete	The type of ticket used for the flight	A categorical one-hot encoded variable for each possible ticket type.	Categorical	No

Feature relevance analysis

After their definition, all the proposed features have been tested for their relation with the target variable (whether the observed trip is a business trip) using two well-known feature selection filter metrics: f-test and mutual information. The former computes the Fischer score between each feature individually and the target, while the latter is based on the mutual information shared between target and feature. This means that the f-test value will capture more linear relationships whilst mutual information will measure non-linear ones.

Table 5.53: F-test and mutual information values for each feature in the training set

	F-test		Mutual Information
	Metric	p-value	metric
acompt2	30.28	~0	0
age_E0	102.24	~0	0
age_E1	48.71	~0	0
age_E2	25.93	~0	0.00023
age_E3	182.15	~0	0
alo_Apartamento/Viv.alquiler	51.23	~0	0
alo_Hotel	128.49	~0	0.013
alo_Lugardetrabajo	172.33	~0	0
alo_Otros	0.45	0.5	0
alo_Segundaresidencia	11.71	0.0006	0
alo_VienedeResidencia	2.21	0.14	0.0039
alo_Viviendafamilia/amigos	219.49	~0	0.012
approximate_age	76.79	~0	0.019
arrival_hour	1.09	0.3	0.0005
arrival_weekday	197.56	~0	0.10
days_rel_festivity	64.08	~0	0
departure_weekday	4.85	0.03	0.016
edu_4	0.84	0.36	0
edu_Básicos	48.16	~0	0
edu_Secundarios	385.63	~0	0
edu_Superiores	451.31	~0	0.009
gender	406.99	~0	0.023
mod_AutobúsCortesía	6.60	0.01	0
mod_AutobúsInterurbano	1.20	0.27	0
mod_AutobúsLargaDistancia	68.38	~0	0
mod_AutobúsPúblico	0.37	0.54	0
mod_AutobúsUrbano(EMT)	36.64	~0	0

	F-test		Mutual Information
mod_Cercanías	5.85	0.02	0
mod_CocheAlquiler	5.32	0.02	0
mod_CochePrivadoAcompañante	115.53	~0	0.0004
mod_CochePrivadoConducidoPax	21.81	~0	0
mod_Metro	112.49	~0	0.0039
mod_Otros	0.04	0.85	0
mod_Taxi	459.31	~0	0.032
mod_Uber/Cabify	17.24	~0	0
month_11	12.77	0.0004	0.0007
month_6	12.77	0.0004	0
nios2	39.88	~0	0
npers	64.70	~0	0.11
res_Europe	1.51	0.22	0.0018
res_Latinamerica	148.11	~0	0
res_Rest	17.07	~0	0
res_Spain	77.17	~0	0.0025
taus	201.32	~0	0.25
tick_Combinado(paquete)	9.65	0.002	0
tick_Otras	0.53	0.47	0
tick_Preferente	45.32	~0	0
tick_Promocional+Turistabásica	23.10	~0	0
tick_Turistaconextras	9.09	0.0026	0
working	1,340.55	~0	0.4

It should be noted that these metrics are mainly for feature comparison purposes and therefore the comparison should be focused on the relative differences of different features for the same metric.

Both metrics indicate that the most relevant variable is “working”, which is a binary variable indicating whether the traveller is employed or in any other situation (student, unemployed, retired, etc.). Then, F-test suggests that Taxi mode (mod_Taxi), having high education (edu_Superiores) and gender (gender) are also relevant features in terms of their relationships with the target variable. Mutual information gives more relevance to the number of people travelling together (npers), the day of the week of the arrival date (arrival_weekday) and the length of the stay (taus).

In both cases, all these variables score high with respect to others and only some variation on the top positions is observed. Regarding less relevant features, it can be observed that one-hot encoding variables are the most frequent types of variables. Most of the “other” categorical variables tend to have small relevance for the model in terms of both metrics. It is also worth noting that mutual information is zero for a large proportion of features (30 out of 50), including some features that still have high relevance for the f-test metric.

Analysis

The proposed model consists in a machine learning classifier that takes as input a set of features derived from the EMMA survey in Madrid-Barajas Airport and predicts whether the trip detailed by each respondent has a business purpose or any other. The main hypothesis of this model is that the features available at the EMMA survey (trip details, socio-demographic information, etc.) are relevant to determine whether the passenger travels for business.

Aiming at model understanding, a decision tree classifier has been selected. The decision tree classifier training is based on the evaluation of the relevance of each input variable according to some criteria, typically Gini index or entropy, to select the most relevant variable and create a rule based on that variable. This way, the training process consists in recursively evaluating and splitting data to create new rules that separate data into compliant and non-compliant datasets that will in turn be evaluated to create new rules. When useful data splits can no longer be computed, the algorithm stops in a leaf node. The class prediction is found by following the tree rules up to a leaf node and assigning the class that is more frequent in the resulting leaf node. At the end of the training process, the tree will have developed a set of hierarchically organised rules that performs predictions following the appropriate branch of the tree. This way, the algorithm performs simple non-linear predictions that can be easily understood by visualising the resulting tree.

The dataset has been separated into training and testing data sets, having the testing dataset 33% of the available data. In the training process, 5-fold cross-validation hyper-parameter tuning has been used in order to maximise the performance of the resulting model.

Finally, a random forest classifier algorithm has also been explored for performance improvement. A random forest classifier is an ensemble of tree classifiers that are computed using a random subset of data points and features. While the introduced randomness tends to improve single tree results, it makes the algorithm no longer interpretable.

Model performance will be evaluated in terms of four different metrics:

- **Accuracy**: Number of elements correctly classified overall
- **Precision**: Number of elements correctly classified with respect to the total elements the model predicts to the positive class (is business trip).
- **Recall**: Number of elements correctly classified to the positive class out of the total elements that are labelled with the positive class.
- **F1-score**: Harmonic average of precision of recall. Using the harmonic mean avoids hindering underperformance from one of the two metrics (for instance, it prevents the model to score 0.5 when the precision is 1 and recall 0).

5.2.2. Experiments

After data preparation and feature selection analysis, a battery of experiments has been defined to better cover the use of survey data in the prediction of the business purpose of a trip. These experiments have been conceived to maximise the accuracy of the model as well as to include variables relevant for different applications. Hence, the experiments are performed over data taking into account two different contexts:

- **EMMA features**: train a model to be used over all available features of the given EMMA. The model has access to all possible information and can be then used to infer trip purpose from other surveys or trip information sources that do not include trip purpose.
- **Mobile network data features**: train a model using the data from EMMA but restricting features to only those available in a context of the indicators derived from mobile network registers. This way, the labelled data available through the survey can be used as training data and, once validated, the model can be used in the prediction of the trip purpose for the trips observed in the mobile network data.

Models have been trained based on these two feature sets in order to assess what is the potential for this model in absolute terms and in a data fusion context. Decision trees and random forest algorithms will be evaluated to provide an interpretable classification algorithm and boost performance respectively. To better understand each model's capabilities, their results will be compared to a baseline model based on the following heuristic: consider business trips all those trips:

- where the outward and return trips happen on weekdays of the same week;
- where the outward trip occurs on Sunday and the return is before immediate next Thursday;
- where the outward trip happens on a weekday and the return is on Saturday before lunch.

The trees trained as a result of the decision tree algorithm have also been depicted for further understanding and inspection of the resulting model.

5.3. Validation plan

5.3.1. Overview

The calibrated model has been subsequently validated to ensure that the model produces sensible results and that the resulting model is not overfitted.

5.3.2. Objectives

By analysing long time series combined with airport passenger surveys and using machine learning techniques, a decision tree has been calibrated to identify the trip purpose in long-distance trips, be it business or leisure. In subsequent project stages, this information will help better characterise the modal choices associated to travel decisions.

5.3.3. Validation approach

The validation approach is summarised below:

- Divide the used data into training and test data.
- Stratify the training data into different subgroups using the same cross-validation procedure explained for the passenger characterisation, in order to select the best model through the calculation of the validation score.

- Select the best model and its best parameters among the machine learning models as the one with the highest average validation F1-Score.
- Use the selected model with the best parameters to train a final model on the whole initial training set.
- Validate the resulting model with the test data.

5.3.4. Data and software inputs

To calibrate and validate the long-distance trip purpose, the EMMA surveys have been used. To clean and filter these surveys, the following steps have been taken:

- whenever the survey had missing data in a relevant feature, the survey has been filtered out;
- for simplicity, connection flights have been filtered out;
- categorical variables with more than two categories have been converted to one-hot encoding;
- categorical variables with only two categories have been encoded using a single binary variable;
- dates have been transformed in their day of the week (0: Monday to 6: Sunday)

Regarding the parameters used to validate, the validation inputs are the following:

- training test split: 66/33;
- cross-validation split: 5 folds.

5.4. Results

Table 5.54 summarises the performance metrics obtained for each of the models on the testing dataset.

Table 5.54: Performance of the evaluated models

Model base	Algorithm	Accuracy	F-score	Precision	Recall
Baseline	Heuristic	0.810	0.600	0.528	0.694
EMMA	Tree	0.787	0.667	0.790	0.576
Mobile phone	Tree	0.792	0.605	0.594	0.616
EMMA	Random Forest	0.834	0.721	0.797	0.657
Mobile phone	Random Forest	0.808	0.626	0.600	0.657

The table provides good performance for most of the models, being slightly better those models based on the entire EMMA than the models constrained by the availability of mobile phone data. In general, random forest models show better performance than their tree-based counterparts, even though the increase is not very large.

The baseline model does show good accuracy, close to the one obtained by machine learning models, but the latter are still better, especially in terms of f-score and precision. It is worth recalling that even though the accuracy of the baseline model may be higher than the accuracy of some

models, the available data is unbalanced, being the business trip class the minority class and, therefore, differences in terms of f-score are relevant.

Figure 1 depicts the tree rules learned from the EMMA tree, optimised to a maximum depth of 8.

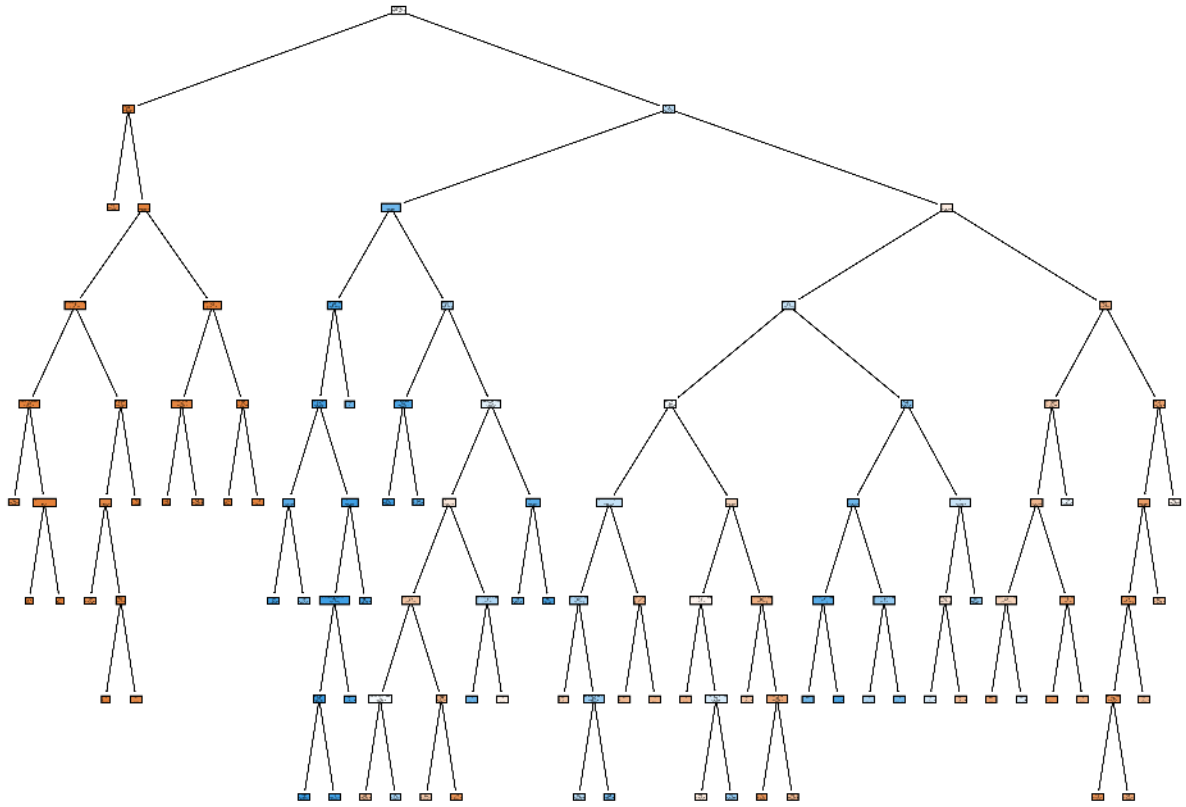


Figure 5.1. Decision tree trained for the EMMA subset variables

The tree takes decision according to the majority class available at each node. Those decisions are represented in blue for the business class and in orange for the leisure (non-business) class. The root node of the tree is the variable “working” that clearly separates data into two sub-trees: the left sub-tree, which accounts for 25% of the entire training set and corresponds to leisure labels in up to 97.5% of those samples, and the right sub-tree, which corresponds to the remaining 75% of the data and is split nearly 40-60% between leisure and business. At this level, the length of the stay feature is considered in both sub-trees.

In order to obtain a more accessible view of the tree, Figure 2 provides the same-model tree but limited in depth to 5. This algorithm is easier to follow but yields a slightly worse performance (0.64 in terms of f-score)

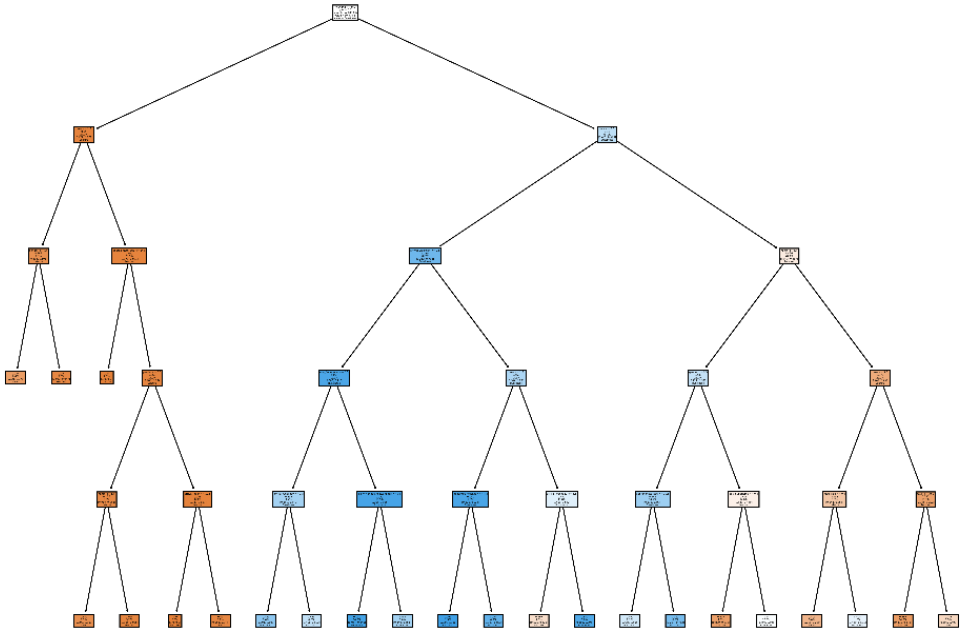


Figure 5.2. Decision tree trained with all EMMA variables but with limited tree depth

Similarly, Figure 3 depicts the resulting decision tree structure when the variables used are restricted to those that can be obtained from mobile phone data.

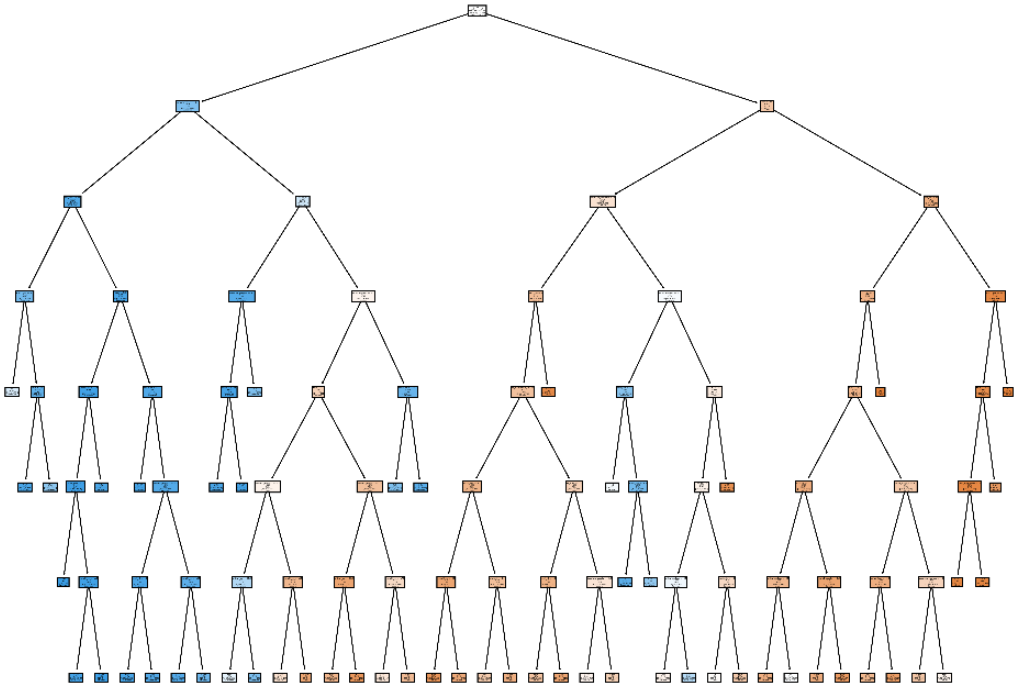


Figure 5.3. Decision tree trained with the EMMA variables restricted to variables that can be derived from mobile phone data and with limited tree depth.

In this case, since the working variable is not used (it is available, but not completely reliable) the root node is based on the taus variable, basically identifying trips shorter than 4 days as business and the rest as leisure. In the left sub-tree, which accounts for 30% of the training data, the amount of business trips is 74%, while in the right sub-tree, with 70% of the total data, business trips account for 35% of them. In this case the business sub-tree uses the arrival weekday variable whereas the leisure sub-tree continues using the taus (trip length) variable.

5.5. Conclusion

In this section we have described the development of a machine learning classifier to determine whether a trip has a business purpose or not. The main data source is airport passenger surveys that ask each passenger to declare their trip details, preferences and other sociodemographic information.

This experiment has led to an extensive analysis of the distinct features that can considered for the model and their relevance for the prediction task. Two datasets have been explored: one that takes into account all possible features to be extracted from the survey and another one that only uses features from the survey that can also be extracted from mobile phone records.

The proposed models are capable of predicting the purpose of the majority of trips, with an F-score of more than 0.72 for random forest models in the case of all possible EMMA features and slightly worse, 0.62, for the case constrained to features available from mobile phone data.

In addition, the following conclusions can be extracted from the feature analysis:

- Some variables present in both data sources, such as the length of the stay, the day of arrival and the gender of the traveller are relevant for the analysis.
- There are variables like “working” that are relevant for the model and can be extracted from mobile phone data, which indicates that the development of these variables is very important for the performance of this model.
- Surveys still provide more varied information, but are much more limited in sample size than mobile phone network data.

6. Airport connectivity and aircraft flows

6.1. Problem statement

The purpose of this study is to extract information on airports that could provide additional information on passengers' door-to-door journey. Mobile phone records may reconstruct the passengers' trip but do not provide the full travel context. An analysis on the transportation supply is relevant to further characterise passenger trips.

Unlike ground transportation, in-vehicle travel time for flights is usually reliable since aircraft are not subject to en-route traffic jams, works or accidents. However, airports, which are connecting points, play a critical part in passenger trips' reliability. In the case of disruptions such as bad weather, Air Traffic Control (ATC) may regulate traffic and differ departure time of flights, causing delays on the scheduled arrival time. This can have a deep impact on passenger connections. Knowledge of an airport situation is essential for passengers to plan a reliable trip. However, even though passengers take these constraints in consideration, disruptive events may occur and cause considerable delays.

In this study, the connectivity and aircraft flows at Paris-Charles-De-Gaulle (CDG) airport are analysed to provide insights on airport "health". Historical scheduled flights data obtained from OAG schedule analyser are used to introduce the connectivity and aircraft flow concepts. OAG schedule analyser [14] is a data source which contains historical airline schedules. This database is updated as soon as an airline declares a schedule change. We collected all scheduled flights at CDG in January and December 2019. Then, an analysis based on historical RDPS (Radar Data Processing System) data around CDG is performed during a disruptive event: the French ATC strike of December 2019.

6.2. Methodology

6.2.1. Evaluation of CDG connectivity

Connectivity analysis in nominal situation

Unimodal analysis

Airport capability to reach several destinations is essential for economic growth. It allows tourism flows, trade of goods, cross-border investment and knowledge exchange. The connectivity of an airport is defined by the International Air Transport Association (IATA) as a measure which reflects the scope of access between a country and the global air transportation network [15]. This connectivity can be measured with different metrics: NetScan Model [16], Weighted indirect connection number [17], Connectivity ratio [18] or Bootsma connectivity [19]. In this study, we only focus on the number of feasible connections and on the "quality" of the connection from the passenger's perspective. The analysis is conducted at Paris-Charles-De-Gaulle (CDG) airport and focuses on domestic flights connecting with international flights. The minimum connection time within Schengen area and outside Schengen area are respectively set at 30min and 45min. A limit of five hours for maximum connection time is set [20]. Indeed, above this time a connection is not considered as attractive for passengers. In this section, the assumption that a connection is feasible between every airline is made. The data used are the OAG flight schedule data of December 2019.

For each day, for each Origin (in France) and Destination (abroad) airports pair, the number of feasible connections is computed. This number is then aggregated by countries. Figure 5.1 displays the median daily number of connections for each Origin-Destination (OD) pair.

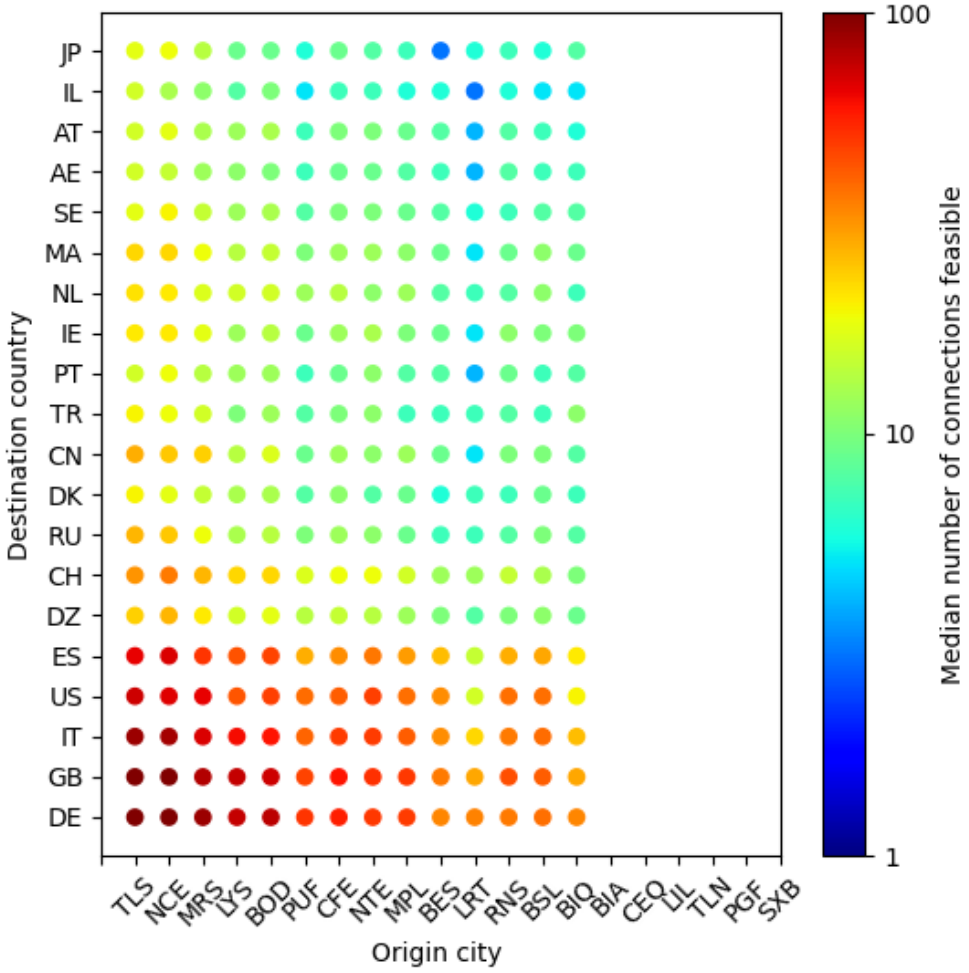


Figure 6.1: Median number of connections per day from France to international through CDG airport for December 2019. Only the 20 most connected countries are displayed.

According to the flight schedules, 110 destination countries were accessible by transiting at CDG airport in December 2019. For visualisation clarity, only the 20 most connected countries are displayed. The best-connected cities in France are Toulouse (TLS), Nice (NCE), Marseille (MRS), Lyon (LYS) and Bordeaux (BOD). They frequently connect Germany (DE), Great Britain (GB), Italy (IT), United-States (US) and Spain (ES). This figure gives insights on OD pair robustness. Indeed, the median value (M_{OD}) can be seen as the expected number of connections each day for an OD pair. The higher this value is, the lower the risk for passengers of failing to reach their destination. Even in the case of delay and a missed connection, passengers could be re-assigned to another flight within the same day. On the contrary, a lower M_{OD} value means less connections each day and a higher risk of not reaching the final destination in case of disruption.

Multimodal connectivity

In this section, incoming flights from France to CDG are replaced with High-Speed Train (HST) connections. Indeed, in France, since March 15th, 2021, short distance flights are forbidden if an alternative by train exists in less than 2h30 [21]. Cities such as Bordeaux, Lyon or Marseille could be deeply impacted by this measure since they can reach Paris-CDG by train in less than three hours. In the same way as in the previous subsection, the computation of the median daily number of connections is made for the rail-flight connection type. The schedule of HST transiting through CDG has been obtained by processing SNCF GTFS data of December 2019. The same OD pairs have been kept. Check-in and security processing time need to be considered for the connection time between a train and a flight. Thus, the minimum connection time is set to 60 min for Schengen connections and 90 min for non-Schengen connections.

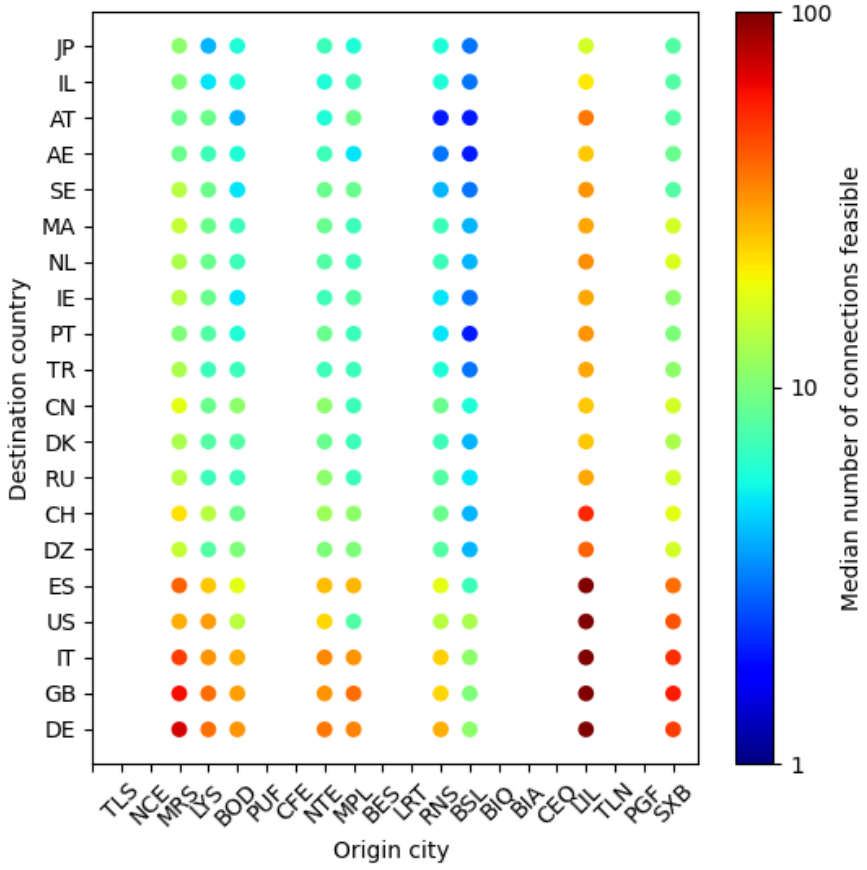


Figure 6.2: Median number of train-flight connections per day from France to international through CDG airport for December 2019. Only the 20 most connected countries are displayed.

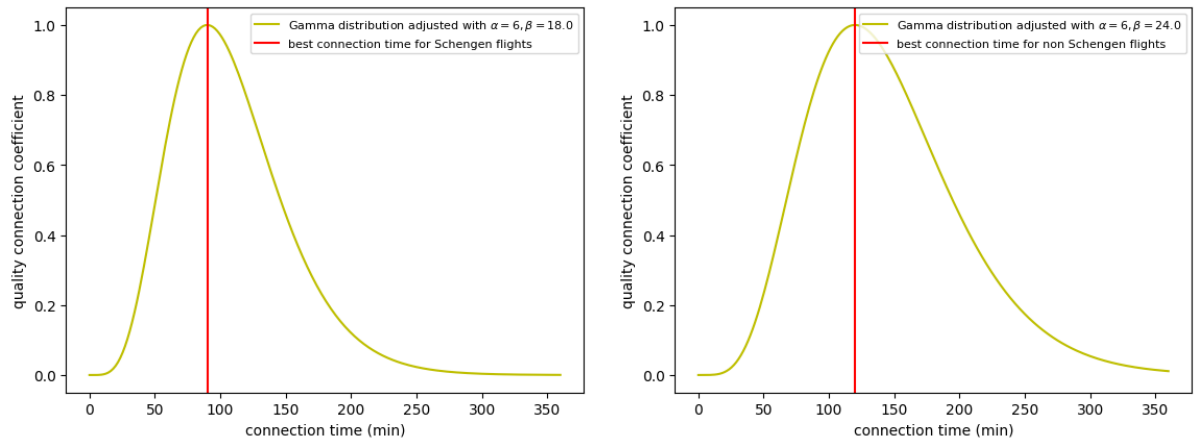
First-hand, cities without HST stations lose their connectivity with CDG. For example, Toulouse, which is one of the most connected cities by flight, is no more connected to CDG by HST. Nice also loses its good connectivity despite it has a HST station. Indeed, trains from Nice to CDG train station were scheduled only on three days during December 2019. Other cities such as Lille (LIL) or Strasbourg (SXB), which were not connected by flights due to their proximity with CDG, are now connected by train. Lille becomes the most connected city with CDG.

Connection score

The feasible connections previously presented do not have the same "value" for passengers. Indeed, passengers expect to travel seamlessly, without neither waiting too long at the airport nor risking missing their connection. In order to evaluate the quality of the connection, a new measure is proposed.

The CDG website [22] suggests allowing at least 70 minutes if passengers transfer within the same terminal and 90 minutes if passengers must change to another terminal. Since most passengers favour a stress-free trip, in case of delay on the arriving flight, the optimal connection time for Schengen area connection is set at 90 minutes, while the optimal connection time for non-Schengen area connection is set at 120 minutes. However, airlines could offer connection times shorter than that. As before, the minimum connection time for domestic (Schengen) and international (non-Schengen) connections is set at 30 min and 45 min, respectively.

A quality coefficient q between 0 and 1 is used to define the quality of a connection. In order to represent this q, a scaled Gamma law is used. A quality coefficient of 1 corresponds to the optimal connection, while a too long or too short connection has a q close to 0. The Gamma laws used for the model are displayed in Figure 6.3.



(a) (b)

Figure 6.3. a) Gamma distribution depicting the quality of a domestic (Schengen) connection with an optimal connection time at 90 min (b) Gamma distribution depicting the quality of an international (non-Schengen) connection with an optimal connection time at 120 min

The non-symmetry of this law makes sense for the study, as passengers would prefer waiting a little bit longer than risking missing their connection. As an example, a connection with *q=0.4* could represent a connection time of 46 min as well as 156 min. Thus, passengers are assumed to value these two connection times the same way. Indeed, having a 46 min connection time is a stressful situation for passengers who risk missing their connection if a delay occurs. Similarly, waiting too much time at the airport could be perceived as a loss of time [23].

These quality coefficients can also be used to quantify the connectivity quality for an OD. Indeed, the volume of feasible connections is as important as the quality of connection time. The connectivity score for one OD could be defined as:

$$S_{OD} = \sum q_c \forall c \in C_{OD}$$

with C_{OD} the set of feasible connections for the OD pair. This connectivity score is a trade-off between the volume of feasible connections and their respective quality q.

Table 6.55 gives an overview of the quality coefficients q for several connection times.

Table 6.55. Connection score for different connection times

Connection Time	30	60	90	120	150	180	210	240	270	300
Score of Schengen connection	0.12	0.7	1	0.8	0.46	0.22	0.09	0.03	0.01	0
Score of non-Schengen connection	0.04	0.38	0.83	1	0.87	0.62	0.39	0.22	0.11	0.05

For example, for an OD in non-Schengen area, the score of a unique 2-hour connection is equivalent to the score of two non-optimal connections of respectively one and three hours: $q_{120} = q_{60} + q_{180}$. Indeed, a two-hour connection fits passengers' expectation in terms of trade-off between the waiting-time and the risk of missing the connection. This trade-off could be also obtained with two connections: if one passenger misses its one-hour connection, he can be re-booked into the second flight but with a longer waiting-time.

6.2.2. Evaluation of CDG aircraft flows

Delay increases with capacity congestion [22]. A high number of operating aircraft within the airport can induce traffic jams on taxiways and conflicts at the gates which can also be translated into flight delays. Thus, tracking the scheduled aircraft flow throughout a day at an airport can help identify when delays are most likely to happen. This information is valuable for passengers in order to plan a reliable trip especially when this trip includes several flight legs or transport modes. The scheduled flights of January 2019 provided by OAG are analysed in this section.

In order to estimate the flow evolution at CDG airport, an event is created for each scheduled flight. The event consists in a tuple of size 2. The first element describes whether it is a departure or an arrival flight. The second element is the time when the aircraft arrives at the airport for arrival flights or when the aircraft leaves the airport for departure flights. Then, all the events are sorted by time and a counter is implemented in order to track the aircraft flow evolution. Since we do not know how many aircraft are parked at the airport during the night, the value of the counter is initialised to 0. Thus, the number of aircraft can be negative if for a given time more departure events happened than arrival events. Even if the true number of aircraft at the airport is unknown, this counter provides a simple way to track the aircraft flow dynamics and helps identify patterns of delay occurrences. Figure 6.4 and Figure 6.5 respectively illustrate the scheduled aircraft flow at CDG during the first week and on Fridays of January 2019.

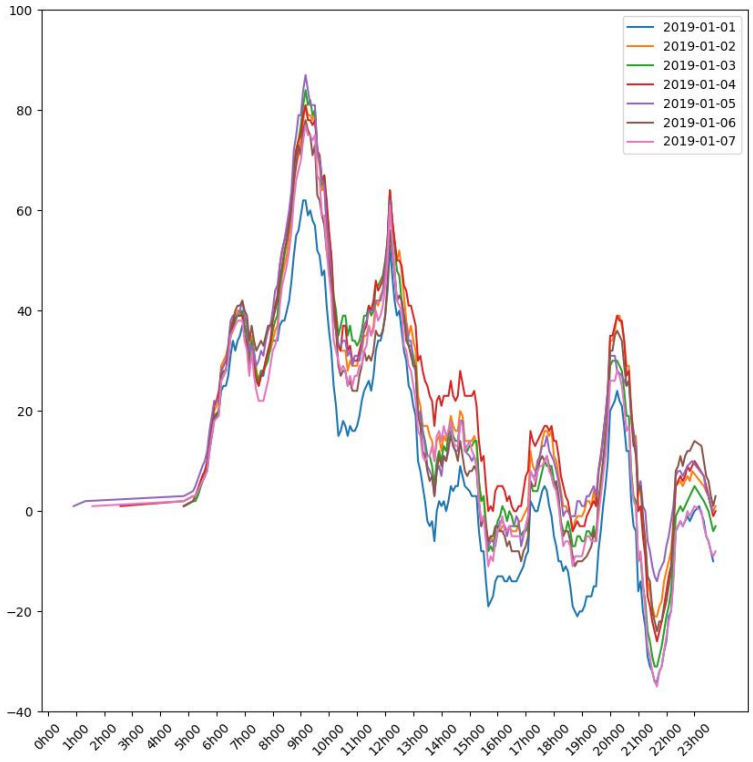


Figure 6.4 Evolution of scheduled aircraft flow at CDG during the first week of January 2019

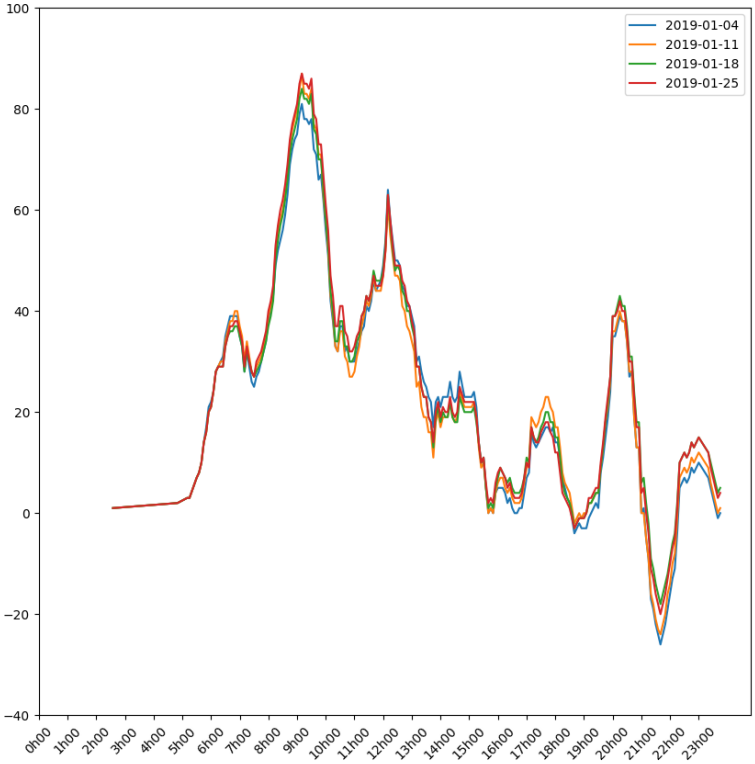


Figure 6.5: Evolution of scheduled aircraft flow at CDG on Fridays of January 2019

Each curve represents the scheduled flow of one day. For each of them three peaks are identified respectively around 9am, 12am and 8pm. These peaks are characterised by a high increase which corresponds to a higher throughput of scheduled arrival flights than departure flights. For simplicity, we associate this increase as a flow of scheduled arrival flights at the airport. This increase is followed by a drop of the number of aircraft which similarly can be seen as a flow of departure flights. These peaks identify the periods of the day where the largest number of aircraft are expected within the airport coupled with a high frequency of aircraft movements. These are critical for the airport because runways, taxiways and gates are likely to be congested, which can induce delays, especially for departure flights.

One can notice in Figure 6.4 that the global flow evolution for January 1st, 2019 is lower than for the three other days. The total number of departure and arrival flights is however like the other days. One possible explanation is that this day is a public holiday and thus less congestion is expected at the airport through a flow of arrival and departure flights more scattered during the day.

Airlines are often planning their schedules on a weekly basis as explained in [24]. The reliability of planning a trip on a Tuesday or on a Friday is likely to differ. In order to confirm this assumption, a principal component analysis (PCA) is applied to the flow evolution data. This method helps to compare a set of data points by extracting directions that best fit the data and thus allows for a reduction of the dimension space. Here each operational day is represented by a vector of dimension 288, which represents the relative number of aircraft within the airport every five minutes of the day. Only the two first components are kept after PCA to plot each day as a point in a 2D graph. An isolation forest is then applied in order to detect outliers with a contamination threshold at 0.1. This means that if the number of operational days is equal to 100, 10 days will be detected as outliers. These outliers can refer to non-nominal situations which can require further analysis in order to anticipate abnormal aircraft flows. This can help to reduce delay propagation and thus to design a more reliable schedule. Figure 6.6 illustrates the result of the PCA and the Isolation forest done on flights scheduled at CDG in January 2019.

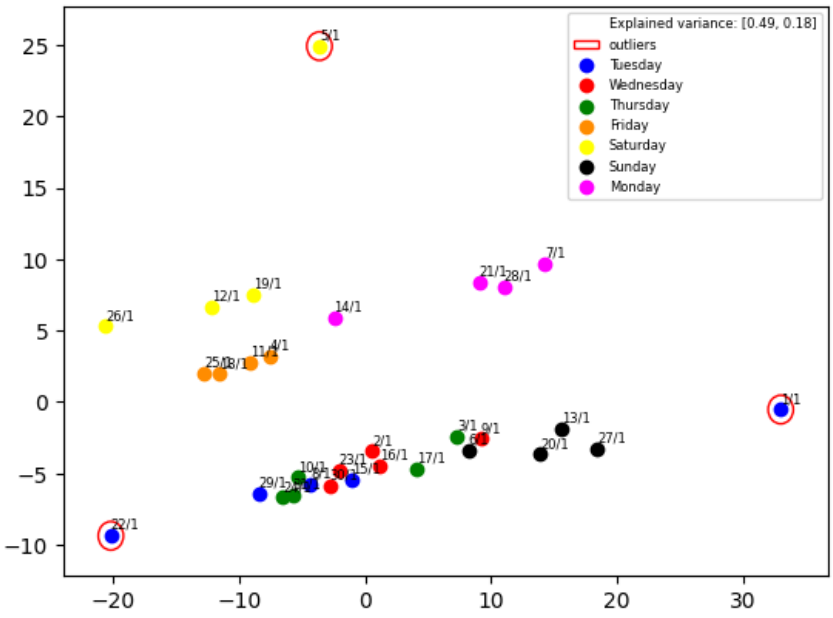


Figure 6.6: PCA applied on scheduled aircraft flow at CDG in January 2019

In this Figure, each day of the week is represented by a colour and outliers are circled in red. Two first principal components explain 67% of the total variance. One can notice that points are clustered by day of the week. This means that the aircraft flow dynamics for two Mondays or two Fridays is likely to be similar, which confirms the previous assumption that airlines are planning their scheduled on a weekly basis. However, Tuesdays are more spread than the other days; further analysis would be useful in order to understand this phenomenon.

Regarding the Isolation Forest, January 1st is detected as an outlier. This supports the previous analysis done on Figure 6.4. Public holidays seem different in terms of aircraft flow patterns and further analysis on delay propagation and/or missed connections would be relevant in order to study reliability of trips during these days.

The yellow jackets movement is also a possible explanation for the detection of 5th of January as an outlier. Indeed, this movement influenced the tourism in Paris [26], with a 6.8% reduction on flight tickets booking for the beginning of 2019. Several major demonstrations were planned on January 5th according to act VIII of the yellow jacket's movement. These demonstrations can explain why this day is categorised as an outlier. Regarding the last outlier, snowfalls were expected on January 22nd, 2019. This event was likely to generate last minute changes on airline schedules and thus abnormal aircraft flow dynamics.

6.3. Case study: ATC strikes December 2019

A nationwide strike occurred in France on December 5th, 2019. French air traffic controllers were involved in this movement, leading to a massive disruption on the air transportation system. In this section, the connectivity and aircraft flows at CDG airport during this event are analysed. RDPS data of the following days were extracted:

- December 4th, 2019: nominal situation
- December 5th, 2019: ATC strike
- December 6th, 2019: recovery day.

These data contain the real arrival and departure time of every flight at CDG. Since these files contain traffic information around CDG, freight or military aircraft are also included, while OAG only provides schedules of passenger flights.

6.3.1. Data pre-processing

A first pre-processing has been done in order to remove non-passenger flights included in the RDPS data. Nine non-passenger airlines have been identified. For instance, flights with a callsign beginning with FDX are operated by Fedex which is a cargo carrier. For the connectivity analysis, flights operated by these airlines are removed from the data set.

Table 6.56: List of non-passenger airlines identified and operating at CDG

Callsign	Operator
ABR	ASL Airlines Ireland
FDX	Fedex
FPO	ASL Airlines France
BCS	European Air Transport (Belgium)
XRC	Express Air Cargo
UPS	UPS Airlines
MNB	MNG Airlines
DHX	DHL International Aviation ME
FAF	French Air Force

The following table shows::

- Nb scheduled flights: the number of scheduled flights according to OAG
- Nb actual passenger flights: the number of remaining flights in RDPS data after removing non-passenger airlines identified in Table 6.56
- Nb actual non-passenger flights: number of flights operated by an airline identify in Table 6.56
- Total nb actual flights: Number of flights in the RDPS data file

Table 6.57: Number of scheduled and actual flights at CDG on December 4th, 5th, 6th, 2019

	2019/12/04	2019/12/05	2019/12/06
Nb scheduled flights (OAG)	1,169	1,192	1,245
Nb actual passenger flights (RDPS)	1,216 (+3%)	988 (-17%)	1,173 (-6%)
Nb actual non-passenger flights (RDPS)	112	111	83
Total nb actual flights (RDPS)	1,328	1,099	1,254

On December 4th, there was a difference of 3% between the number of scheduled passenger flights and the actual one. One possible explanation is that several non-passenger flights have not been detected. For example, Air France Cargo is a freight company operating at CDG and its callsign is AFR, which is also Air France's callsign. Thus, no distinction could be made on RDPS data between these types of flights and an uncertainty of around 3% between schedule and RDPS must be considered. On December 5th, the day of strike, at least 17% of the scheduled flights were cancelled during the ATC strike; on December 6th, 6% of the scheduled flights were also cancelled.

Regarding differences between these two data sources, the identification of flights from schedule to RDPS is difficult since callsigns are not always matching. Only few flights have been identified in the two datasets. Thus, in the following sections, only information on flights from RDPS data is considered, without any knowledge of their respective schedule.

6.3.2. Connectivity analysis under disruption

Overall insights

Airport connectivity could be affected when disruptions occur. These disruptions have deep consequences on passengers' journeys, which undergo delays and cancellations.

The total number of feasible connections between OD pairs on December 4th, 5th and 6th are displayed in Figure 6.7, Figure 6.8 and Figure 6.9 respectively

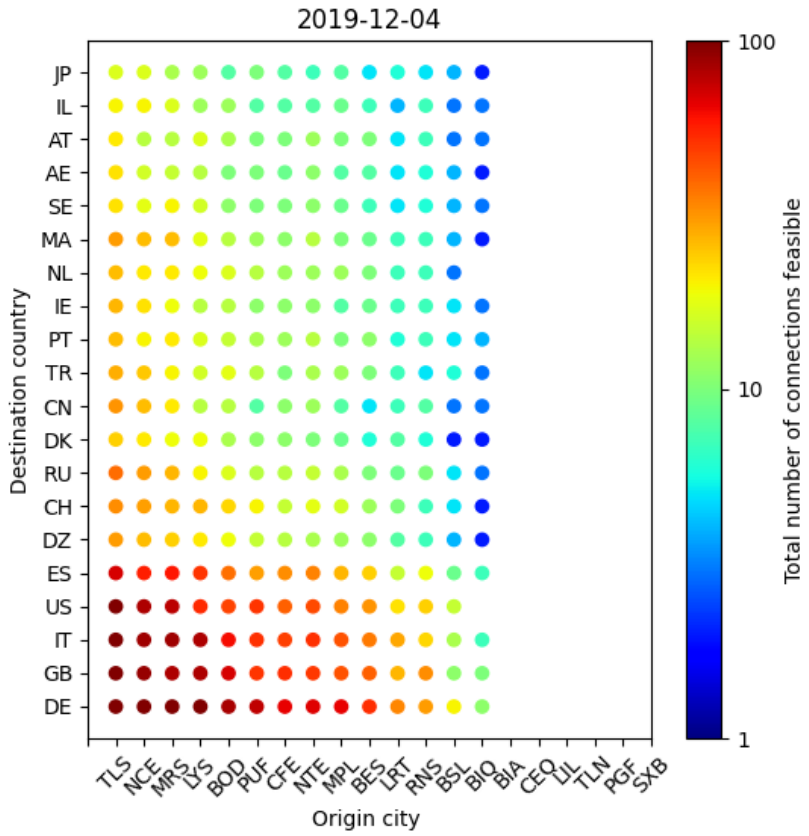


Figure 6.7: Number of feasible connections at CDG on December 4th

December 4th corresponds to a nominal day, when the connectivity is close to the scheduled connectivity illustrated in Figure 6.1. The most connected cities are Toulouse, Nice, Marseille and Lyon. RDPS data fits the planning offered by airlines on nominal days.

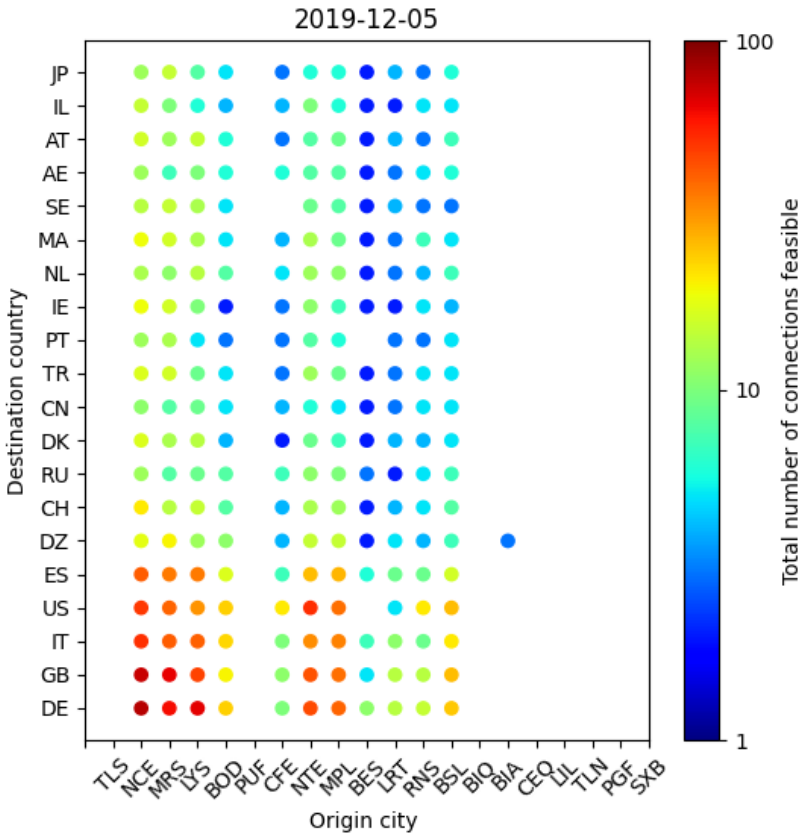


Figure 6.8: Number of feasible connections at CDG on December 5th

On December 5th, Figure 6.8 shows that Toulouse, Biarritz and Pau are not connected anymore with CDG airport. Regarding flights arriving from Toulouse, the only flight lands after 11pm, and therefore no connections are feasible. The ATC strikes deeply impacted these cities and passengers could not reach CDG during the whole day.

Passengers impacted had to adapt their trip to reach CDG from another city or change their transportation mode. Regarding other OD pairs, a decrease can be observed in the number of connections. For example, Brest (BES) made less than five connections with each destination where it makes around ten a day in general. This decrease of connectivity is the consequence of a lot of flight cancellations and delays. Moreover, the decrease in the number of connections reduced the number of alternatives for passengers and a delay in an arrival flight at CDG might have resulted in being stuck at the airport. This situation must be avoided for airlines, which have to accommodate passengers for the night.

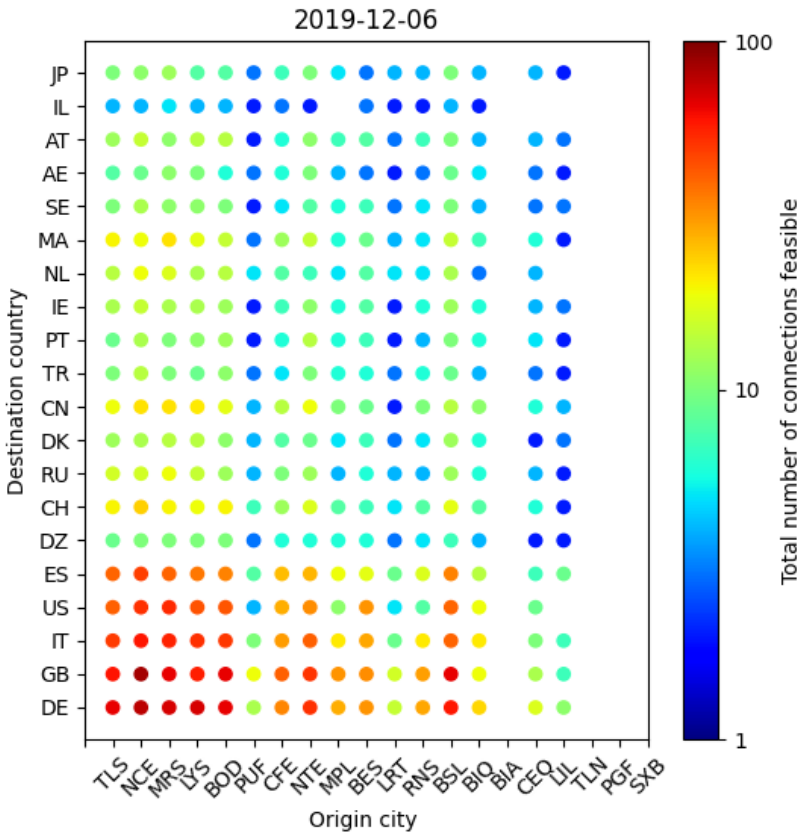


Figure 6.9: Number of feasible connections at CDG on December 6th

On December 6th, the recovery day, most impacted cities during the ATC strike (Toulouse, Biarritz, Pau) started the recovery. However, the connectivity of these cities is still lower than usual.

The number of connections by connection time classes during the three days is shown in Figure 6.10.

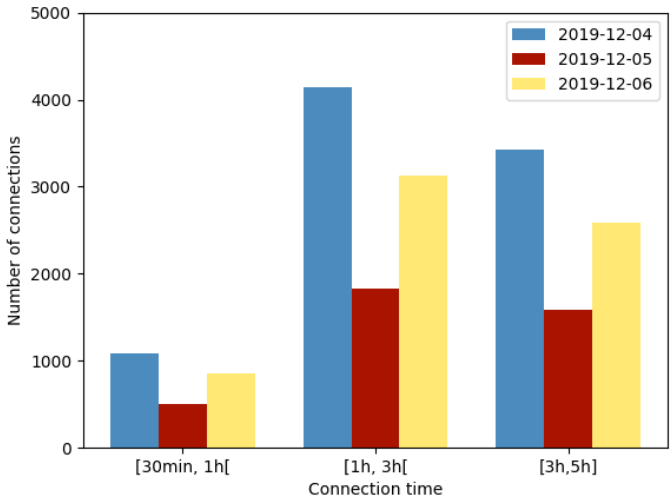


Figure 6.10: Number of feasible connections during the ATC strike event. Connections are divided in 3 classes according to their time.

As previously seen, the number of connections is highly reduced during the strike and does not entirely recover on December 6th. Table 6.58 presents the distirbution among classes for each day.

Table 6.58: Number of connections at CDG on December 4th, 5th, 6th

	2019-12-04	2019-12-05	2019-12-06
30min-1h	1,081 (12%)	497 (13%)	860 (13%)
1h-3h	4,140 (48%)	1,832 (47%)	3,125 (48%)
3h-5h	3,427 (40%)	1,592 (40%)	2,585 (39%)
Total	8,648	3,921	6,570

While the volume of connections is reduced, no connection time class is more impacted than the other during the strike. The distribution is stable during the three days of the event.

Benefits of airline collaboration

Until now, all the connections are assumed feasible without airline constraints. Only the connection time between two flights was considered to assess the feasibility of a connection. Now, the inverse case is considered: a connection is supposed to be feasible only if two connected flights are operated by the same airline. In reality, alliances exist between airlines but these two considerations (connections between all airlines or only within the same airline) allow us to set up extreme limits in the number of potential connections.

Figure 6.11 displays the total number of feasible connections on December 4th with this new constraint. Connections are split into three classes corresponding to their connection time. On December 4th,, 88 airlines operated through CDG airports. Since only connections from France are selected, the number of airlines which can make a connection between a French domestic flight and an international flight is significantly reduced. First, the number of connections is always at least twice higher when no distinctions are made on airlines. Furthermore, for a connection time lower than one hour, only one airline succeeds in making feasible connections.

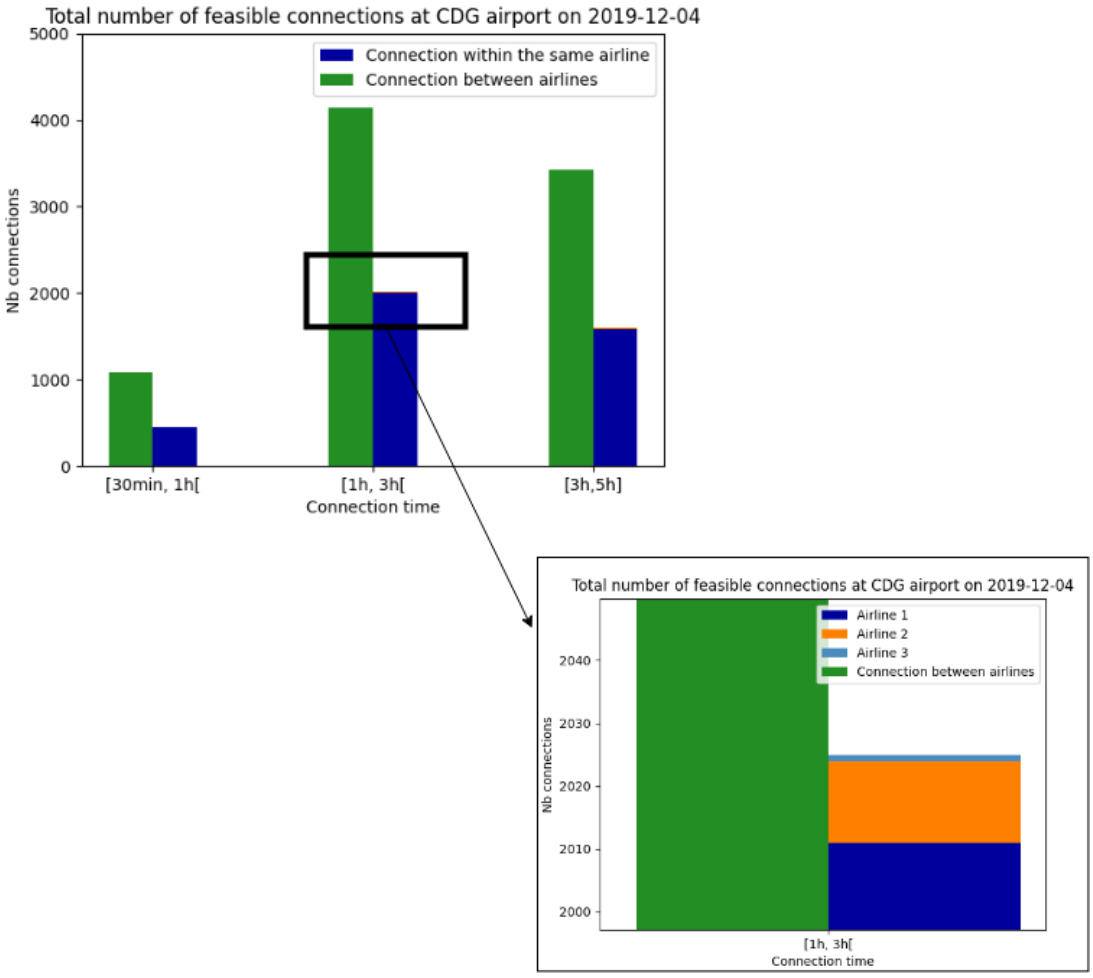


Figure 6.11: Total number of feasible connections at CDG on December 4th. The green bar corresponds to the number of connections possible with no airline distinction. The blues and orange bars correspond to the number of connections feasible within the same airline.

In case of disruptive events such as December 5th, airline restrictions emphasise the decrease in the number of connections. In Figure 6.12 the number of connections is displayed by connection time. As in the previous day, more connections are feasible without airline distinctions. A further analysis on connections within alliances should be made. One could already assert that the number of potential connections will be higher than within a unique airline but still lower than without airline constraints. However, having a snapshot of the actual state of airline collaborations could inform how far the aviation system is from the passengers' "ideal" system where there are no airline constraints. Indeed, one can assume that in case of major disruptions, passengers are more prone to switch between airlines in order to reach their destination.

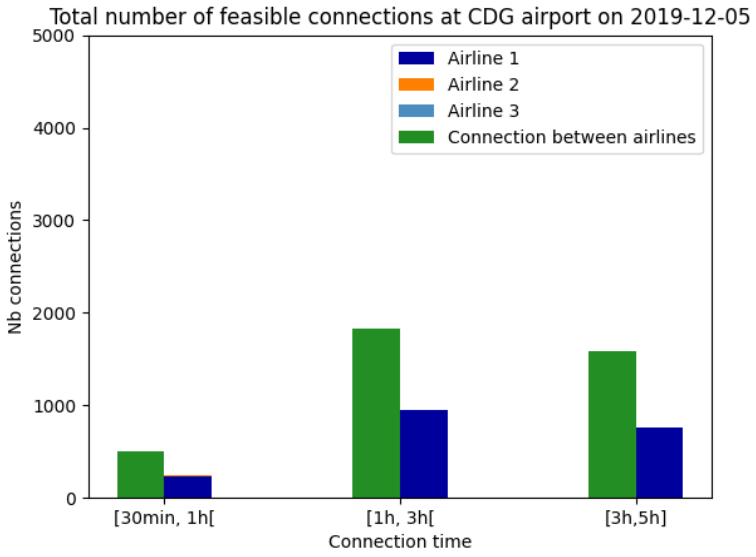


Figure 6.12: Total number of feasible connections at CDG airport on December 5th, 2019

Impact of ATC strike on connection scores

As explained in the first part of the study, the quality score of a unique optimal connection time is equivalent to the score of a set of non-optimal connections. The higher the score is, the better the connectivity between an OD pair. For each day of the analysis period and for each OD pair, the quality score is computed. Figure 6.13, Figure 6.14 and Figure 6.15 present the evolution of this score for each OD pair during the ATC strike compared to the nominal situation of December 4th. For each OD pair and for two days respectively noted day_j and day_i, the quality evolution coefficient is computed as:

$$QE = \frac{S_{OD_j} - S_{OD_i}}{S_{OD_i}}$$

where S_{OD_j} is the quality score of the OD pair on day_j.

- QE represents the evolution of the quality score between day_j and day_i.
- QE = -1 means that no connection remains on day_j compared to day_i.
- QE = 0 means that there is no change in the connectivity between the two days.
- QE = 1 means that day_j has a connectivity twice better than day_i.

Figure 6.13 represents the evolution between December 4th and December 5th. As expected, the connectivity quality decreased on December 5th compared to December 4th. Origins which totally lose their connectivity have a 100% decrease.

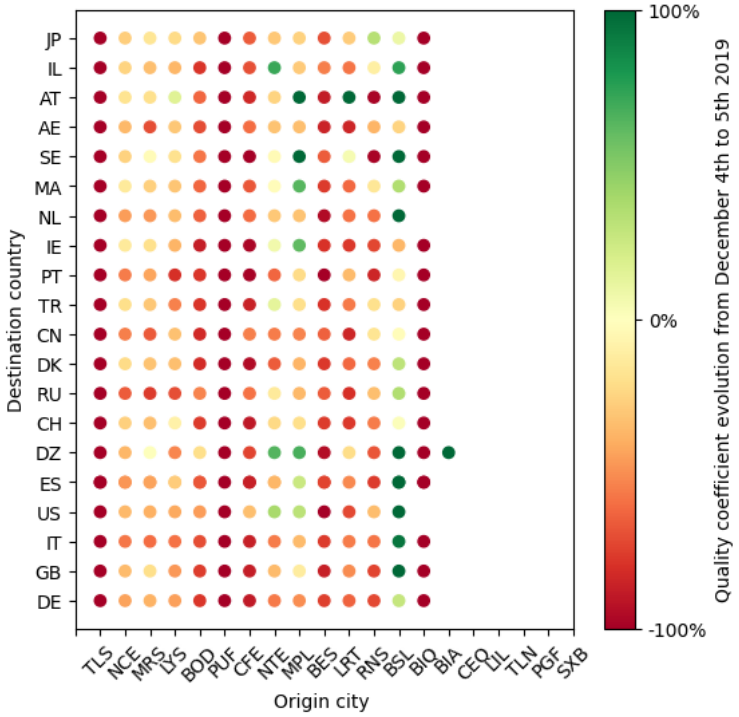


Figure 6.13 : Quality coefficient evolution from December 4th to 5th. Red colour indicates a degradation of the connectivity while green shows an improvement.

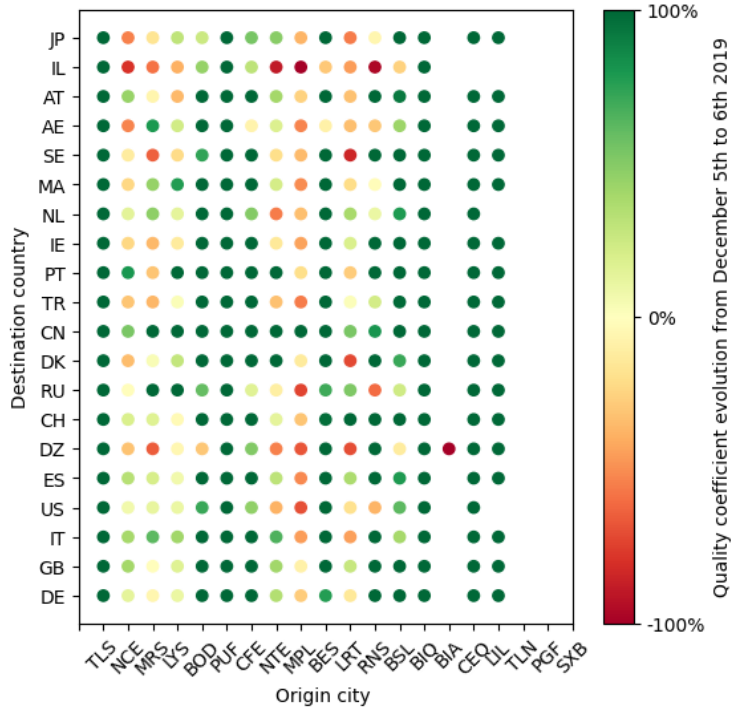


Figure 6.14: Quality coefficient evolution from December 5th to 6th. Red colour indicates a degradation of the connectivity while green shows an improvement.

Figure 6.14 displays the evolution of the quality coefficient between December 5th and December 6th. The connectivity has grown compared to the disruption day which means that the recovery is ongoing. In order to see how well the airport has recovered, the distance to the nominal situation is represented in Figure 6.15.

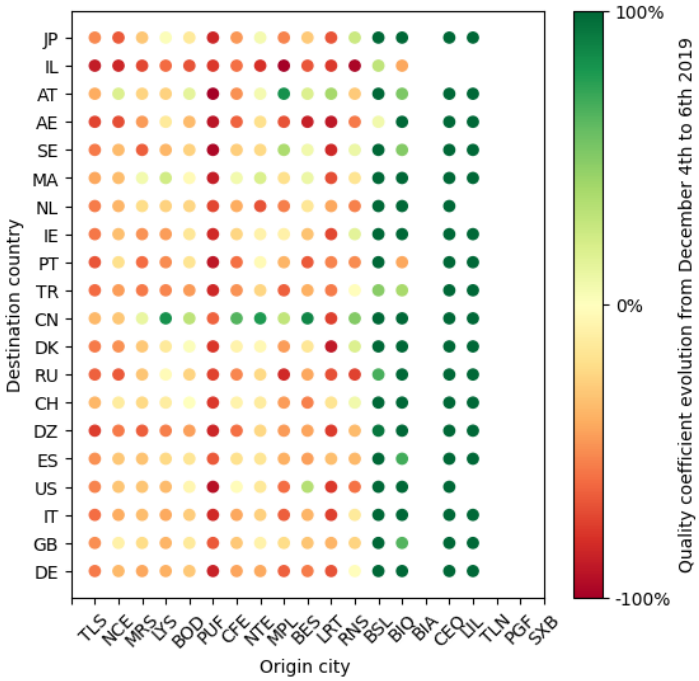


Figure 6.15: Quality coefficient evolution from December 4th to 6th. Red colour indicates a degradation of the connectivity while green shows an improvement.

While the situation on December 6th is better than December 5th, Figure 6.15 confirms that non all the scheduled flights have operated. This is in accordance with results obtained in the pre-processing phase. However, one can notice that new connections are created from Cannes (CEQ) and Lille (LIL) because of arrival flights from these two cities. By looking closer at these flights, we can observe that only one flight from Lille and one flight from Cannes arrived at CDG.

6.3.3. Aircraft flow analysis under disruption

The purpose of this section is to infer average delay at CDG airport only with a macroscopic indicator tracking the aircraft flow at the airport. Indeed, the association between scheduled flights (OAG) and radar flights (RDPS) was not possible due to lack of information. Thus, the delay of each flight was not computable. However, the comparison between the scheduled aircraft flow and the actual one can provide insights on the airport delay status. In this section, non-passenger flights are kept in the RDPS data since they contribute to airport congestion.

Figure 6.16 provides an example of the scheduled aircraft flow compared to the actual one on December 4th, 2019.

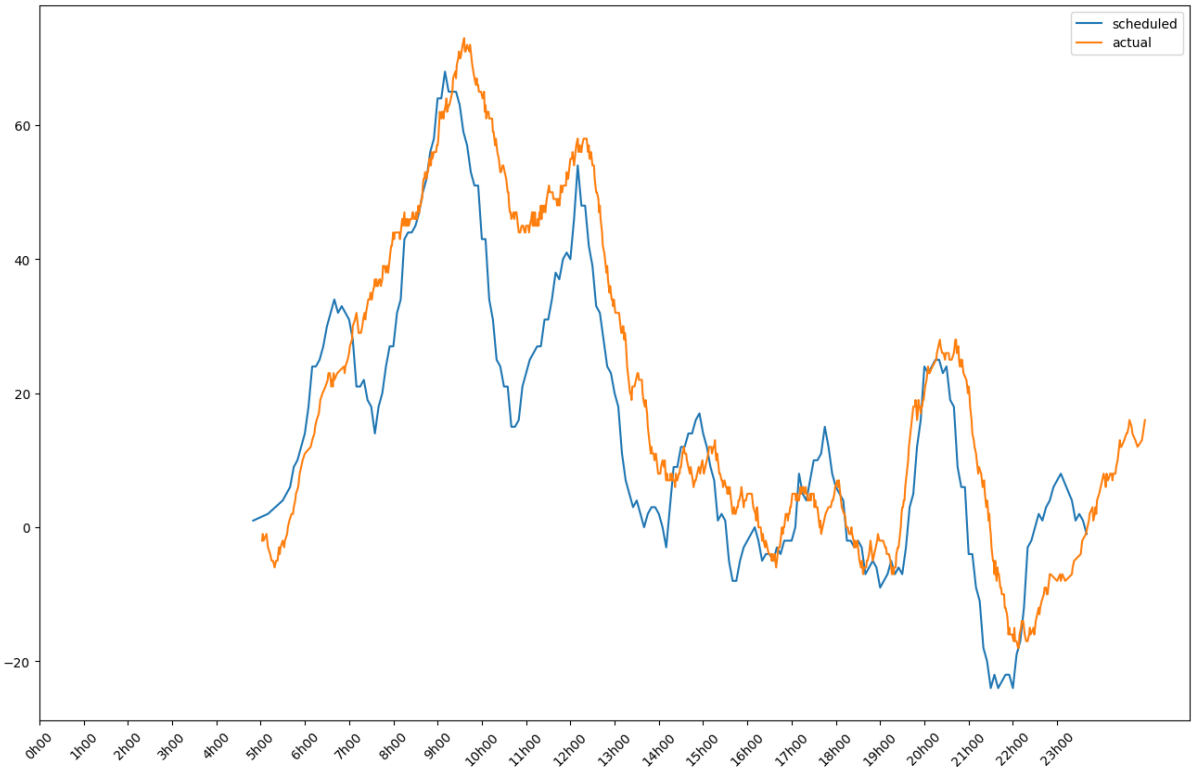


Figure 6.16: Comparison scheduled vs actual aircraft flows on 04.12.2019

First, one can notice that the actual aircraft flow curve seems slightly higher than the scheduled one. One possible explanation is that freight and military aircraft are also seen by the radar but are not included in OAG schedules. Indeed, several non-passenger flights might arrive at the airport in the early morning inducing a higher relative number of aircraft during the overall operation day. Also, small oscillations of the orange curve are noticed. Actually, airlines plan their flights with an accuracy of minutes. Thus, several flights can be scheduled at the same time. However, RDPS data has a temporal granularity of one second and almost all the flights have different actual departure or arrival times. Regarding the signal patterns, the radar aircraft flow dynamics is closed to the predicted one. The actual curve is slightly deviated to the right compared with the schedule one. This can be interpreted as a macroscopic delay accumulated on the airport schedule. Comparing the deviation for each peak (9am, 12am, 8pm) can provide valuable information regarding the delay situation in the airport. These peaks are easily identifiable and can be used as checkpoints to see if the overall schedule is delayed.

In order to automatically identify the three peaks, a smoothing is applied to each curve. This can be done using a Savitzky-Golay filter [27], which smooths the data through a convolution. It sequentially fits the data with a low degree polynomial through the linear least square method.

Figure 6.17, Figure 6.18 and Figure 6.19 present the comparison of the smoothed scheduled aircraft flow and the smoothed actual flow on December 4th, 5th and 6th 2019, respectively.

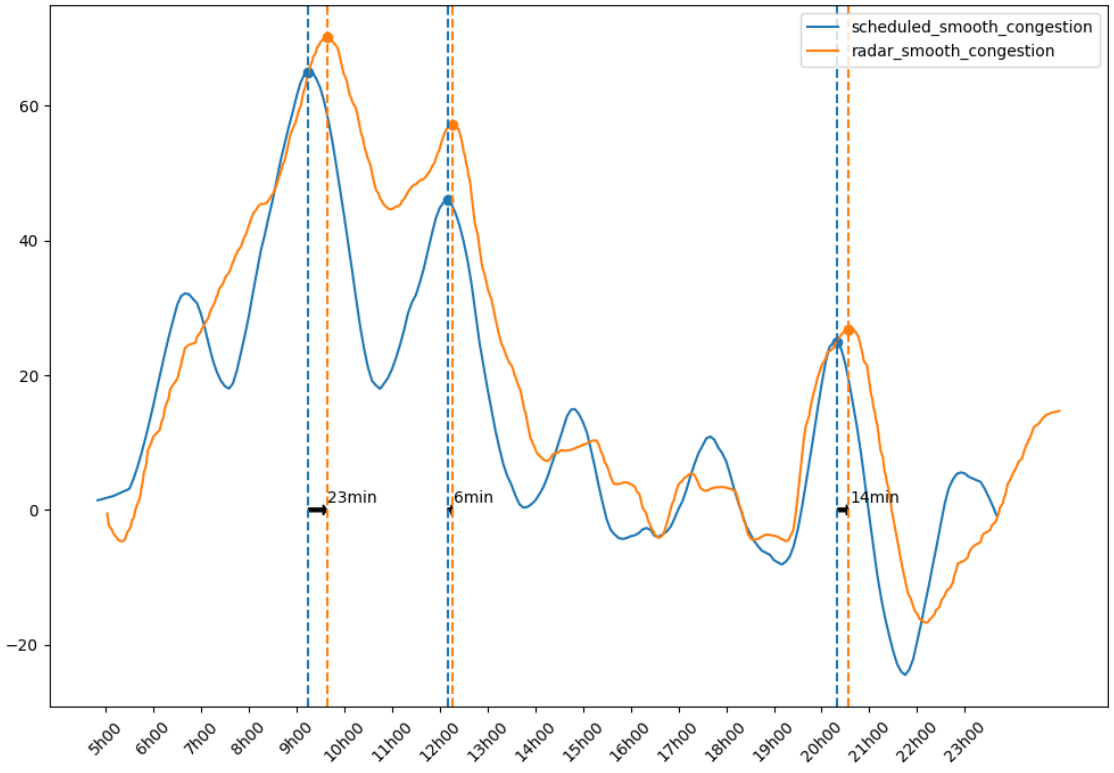


Figure 6.17: Comparison scheduled smooth vs radar smooth aircraft flows on 04.12.2019

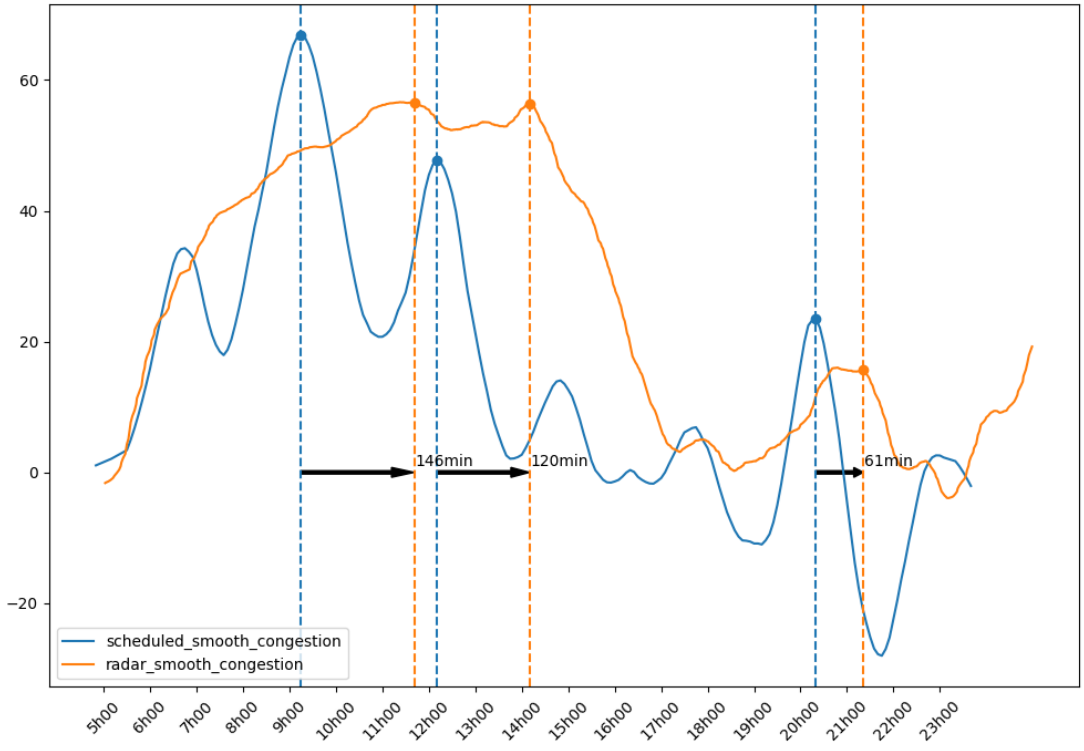


Figure 6.18: Comparison scheduled smooth vs radar smooth aircraft flows on 05.12.2019

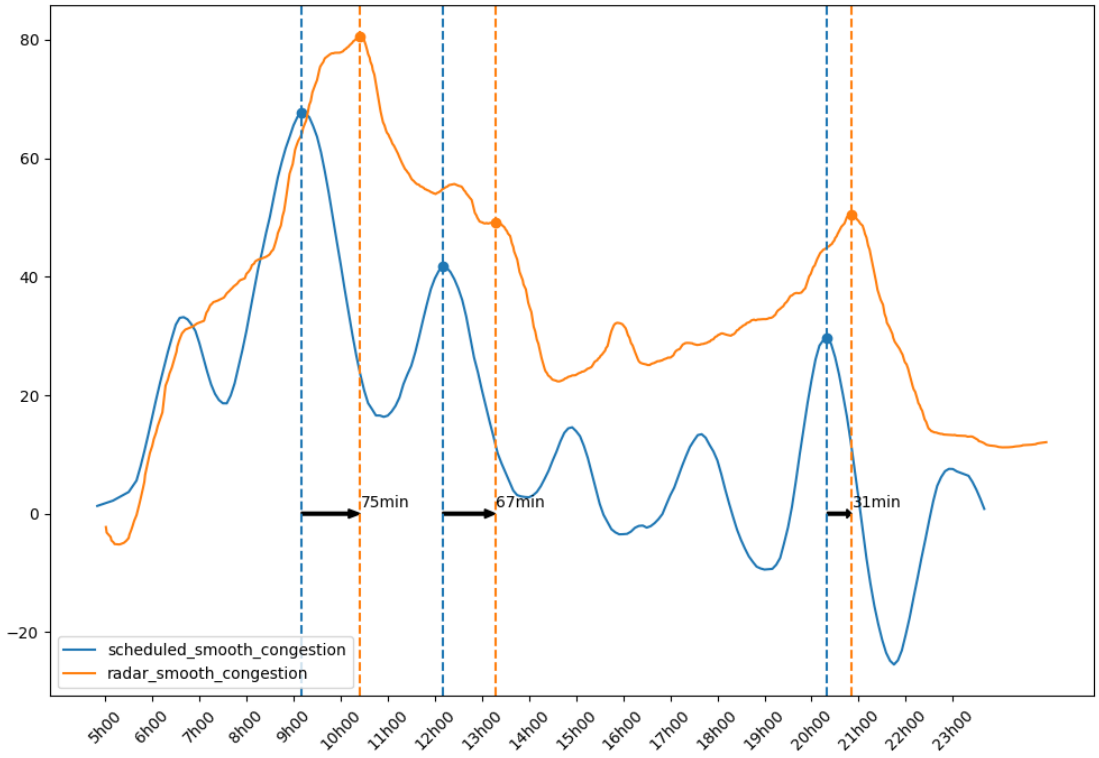


Figure 6.19: Comparison scheduled smooth vs radar smooth aircraft flows on 06.12.2019

On December 4th, the deviation between peaks is respectively equal to 23, 6 and 14 min. This means that a passenger having a flight at 9am is likely to be delayed by 23min on average. Moreover, the flight schedule during this day seems more reliable at midday than at 9 am or 8 pm.

The following day, which corresponds to the French ATC strike, illustrated in Figure 6.18, is clearly a non-nominal situation. The reduction of airspace capacity induced a large deviation between the scheduled and the actual aircraft flows. Regarding the radar plot, peaks are wider. The ascending slope of these peaks, which represent a flow of arrival aircraft, is strongly reduced. This is explained by the airspace capacity reduction. This reduction implies air traffic regulations and thus several arrival flights are postponed in order to space out flights. The delay accumulated by these flights is propagated to the next departure flights operated by the same aircraft. The deviation to the right of the descending slope of each peak, which represents a flow of departure flights, confirms the previous assumption. The 9am peak is deviated by 146 min and the 12am and 8pm ones by 120 and 61 min respectively. Regarding passengers, some of them are likely to be delayed by more than two hours, which corresponds to the optimal connectivity time for non-Schengen flights. Thus, a high number of missed connections can be expected, especially between 9am and 5pm.

December 6th represents a recovery day after the ATC strike and thus the deviation between scheduled and radar aircraft flows is reduced compared to the previous day. However, this deviation is still higher than the nominal situation on December 4th. Delays for the three peaks are respectively equal to 75, 67 and 31 min, which is almost 50 % less than the deviation on December 5th.

One step further is to correlate this aircraft flow evolution to the actual delays of flights. We focused on a case study considering trips from Marseille to New York with a transit at CDG. The association between scheduled flights and actual flights was possible for these trips since callsigns matched between OAG and RDPS data. Thus, the delay of each leg has been computed on December 4th, 5th and 6th 2019. The analysis of these delays is useful to test the insights and assumptions provided in the previous paragraphs.

Marseille has only one airport and three airports are in New York. Two of them are directly connected to CDG during the three days considered. One flight is daily scheduled from CDG to Newark Liberty International Airport (EWR) and seven flights to John F. Kennedy International Airport (JFK). The daily flight schedule remains the same for each day of the study.

Figure 6.20 illustrates the flight schedule at CDG and the potential connections for this trip.

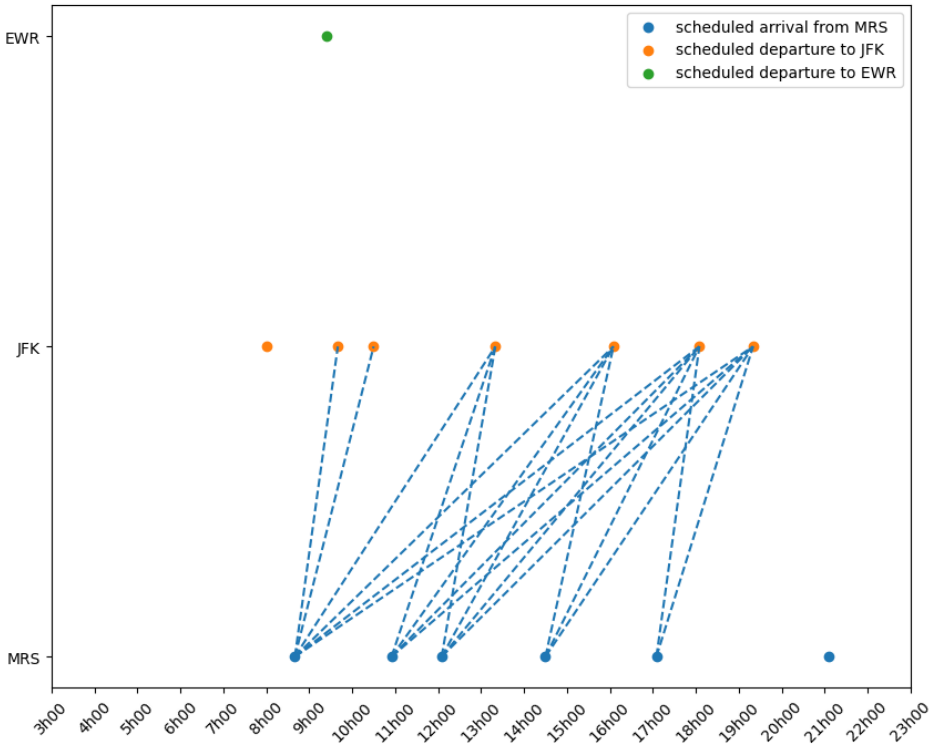


Figure 6.20: Scheduled connection scheme for Marseille-New York trips on 04-06.12.2019 2019

The minimum connection time is set to 45 minutes because it is a non-Schengen connection. One can see that no flight arriving from Marseille can connect with EWR. However, several connections are possible with JFK with a robust connection scheme. For example, let us consider a passenger arriving from Marseille at 08:40 am who missed his connection with the departure flight to JFK scheduled at 09:40 am. This passenger can be re-booked into the flight scheduled at 10:30 am and thus get a delay on its door-to-door trip lower than one hour.

Figure 6.21, Figure 6.22 and Figure 6.23 present the actual connection scheme on December 4th, 5th, and 6th, 2019 respectively.

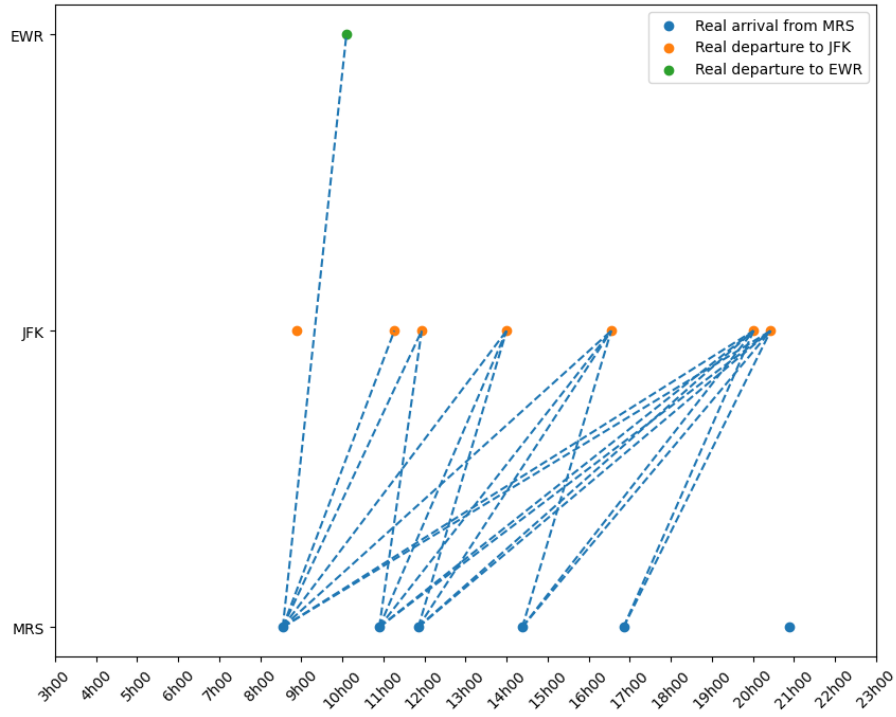


Figure 6.21: Actual connection scheme for Marseille-New York trips on 04.12.2019

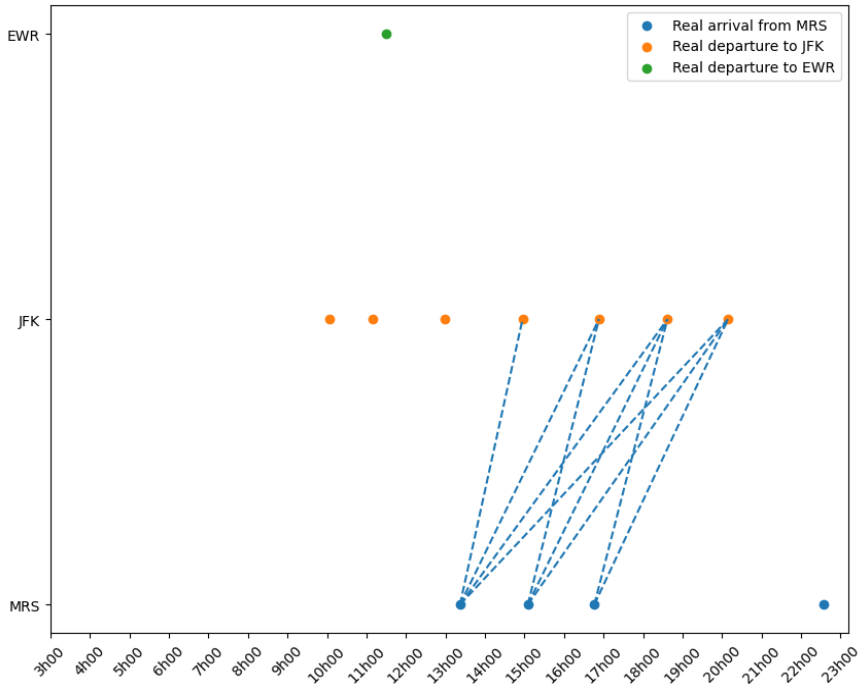


Figure 6.22: Actual connection scheme for Marseille-New York trips on 05.12.2019

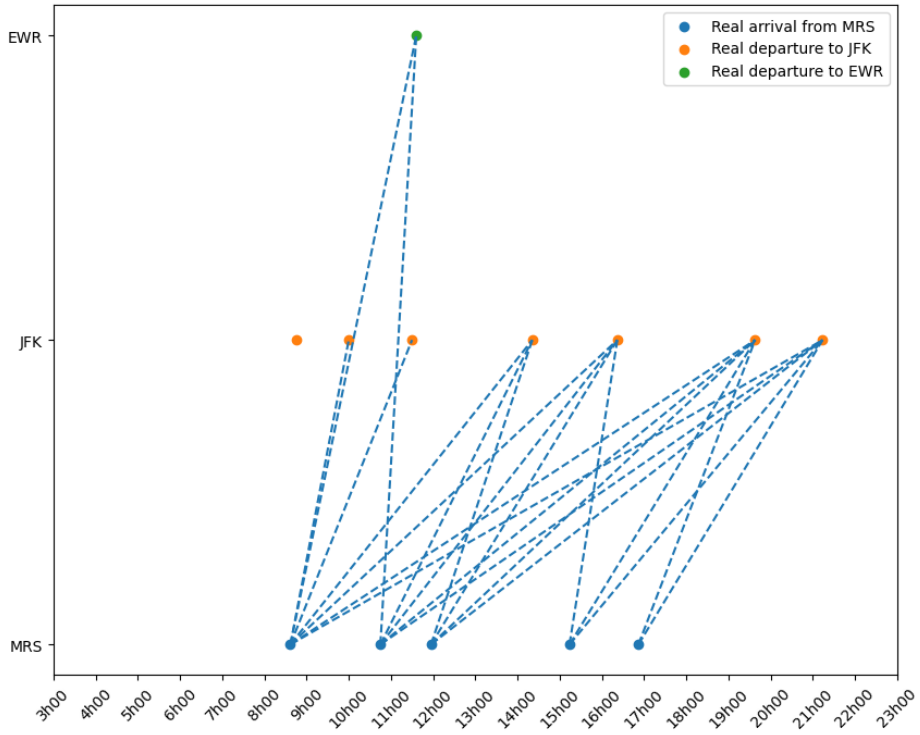


Figure 6.23: Actual connection scheme for Marseille-New York trips on 06.12.2019

All the potential connections defined with the flight schedule are feasible on December 4th. Furthermore, a new connection is now feasible to EWR and another one between the second flight from Marseille and the third flight to JFK. This is due to a delay on both departure flights to New York. The same passenger who planned a connection to the flight to JFK scheduled at 09:40am is impacted by a delay of 90 min due to a departure delay on this flight.

Regarding the day following the ATC strike, the connectivity scheme is depleted with around half of the potential connections not feasible. Only four flights among the six scheduled from Marseille have been operated. The significant delay on arrival flights from Marseille drastically reduced this connection scheme. Departure flights to JFK scheduled at 9:40am and 10:30am are not anymore connectable even if each of them has also been delayed. The same passenger who planned to arrive at 08:40am at CDG from Marseille is now arriving at 01:20pm and will connect at the earliest with a flight to JFK at 03:00pm. Thus, he will be impacted by at least 5 hours and 20 minutes of delay. Finally, the connectivity scheme on December 6th looks like the scheduled scheme, with all the scheduled connections feasible. The last flight from Marseille was cancelled or arrived after midnight but it does not have an impact on the connection scheme. Two arrival flights from Marseille can connect with the flight to EWR. The same passenger will have only a 20 min delay on its schedule due to a departure delay applied to the flight to JFK. However, this delay is likely to be absorbed during the flight to New York.

Figure 6.24 illustrates the cumulative delays for each leg of the Marseille-New York trip. Negative delays are not considered, i.e., earlier arrival and departure flights are considered on-time with a delay equal to 0 min.

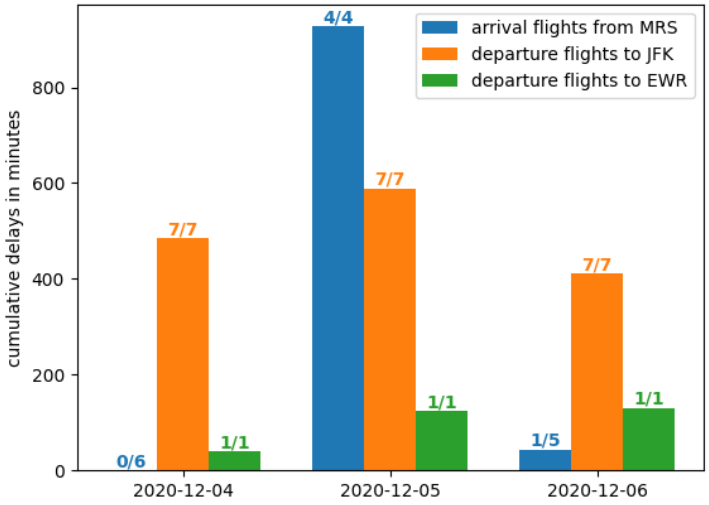


Figure 6.24: Cumulative delay of flights legs included in Marseille-New York on 04-06.12.2019. The ratio above each bar indicates how many flights were delayed.

All flights from MRS to CDG are on time on December 4th. However, all the transatlantic flights are delayed by more than one hour on average. During the ATC strike, the cumulative delay soars for flights from Marseille with almost four hours of delay per flight on average. The recovery can be seen on December 6th, with only one flight delayed from Marseille. Transatlantic flights seem less impacted by the ATC strike. The cumulative delay for flights to JFK is slightly increased by 15 min on December 5th. The only flight going to EWR is delayed every day. The delay on December 4th is equal to 40 minutes and is multiplied by two on the ATC strike day and the day after. Figure 6.25 shows the delay distribution of the considered flights during the ATC strike.

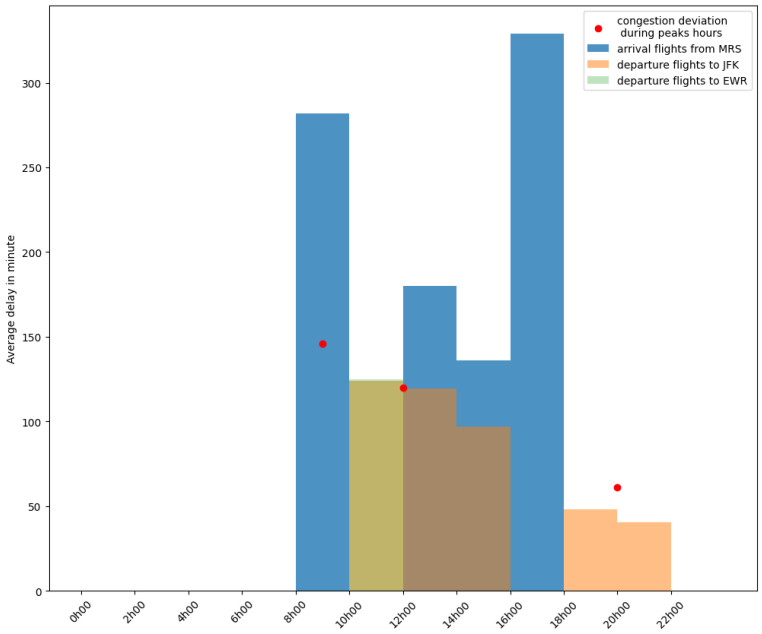


Figure 6.25: Delay evolution of flight legs included in Marseille-New York trips on 05.12.2019

Each colour is associated with a flight leg and the red points show the estimated delay provided by the aircraft flow analysis described in subsection 6.2.2. We can notice that the average departure delay of transatlantic flights is closed to the predicted one at 12am and 8pm. However, the delay of the arrival flight from Marseille during the 9am peak is twice higher than the estimated one. This estimated delay provides insight on the average airport delay status and thus focusing on only three legs is not enough to get satisfying results. Collecting more data on delayed flights is required in order to confirm the insights provided in subsection 6.2.2.

6.4. Conclusions

Airports are crucial connecting points in the door-to-door passenger trip. A unimodal and a multimodal connection schemes at CDG have been drawn. These schemes highlight which French cities are most connected to the world and the potential benefit of considering a multimodal trip for cities having a railway station directly connected with CDG. They also provide an insight on the robustness of OD connections. A quality score based on the connection time of each potential connection has been introduced to evaluate the connectivity at an airport from the passenger's perspective.

The analysis of CDG's aircraft flows makes it possible to identify three peak hours (9am, 12am, 8pm) where passengers are more likely to be delayed. A PCA has been implemented on the scheduled aircraft flow of each day of January 2019 to identify non nominal days. Further analysis would be useful in order to understand the impact on passengers during these days.

Finally, the study of the French Air Traffic Controllers' strike on December 5th, 2019 enables to draw several observations. Firstly, the airport connection scheme is drastically reduced when a disruptive event occurs. This assessment is more evident if we only consider connections within the same airline. The aircraft flow analysis highlights important delays for the passengers and thus a high risk to miss their connections.

Regarding future work, the analysis of real rail data would be useful to draw an actual multimodal connection scheme during disruptive events and see the potential benefits of multimodal coordination. Moreover, knowing the actual overall delay status at the airport would be helpful in order to validate the aircraft flow indicator as a measure of the average airport delay. Knowing the number of connected passengers between two flights would enable the estimation of the real number of missed connections.

7. Conclusions

The present document describes a number of new methodologies and algorithms aimed at improving the characterisation of long-distance travel demand and supply. This enhanced characterisation will be used in the development of the TRANSIT simulation framework to better evaluate the different intermodality concepts proposed in D2.1 according to the performance framework defined in D3.1.

The work on travel demand has focused on enriching the information on door-to-door passenger trips that can be extracted from mobile network data, by improving the characterisation of the passengers' sociodemographic profile, mode choice in airport access, and long-distance trip purposes.

- Improved age and gender segmentation has been derived using observed mobility patterns. In the next project stages, this information will be used to segment passengers and take into account the influence of their sociodemographic characteristics on long-distance travel behaviour (destination choice, mode choice, etc.).
- The resolution of mobile network data makes it difficult to distinguish the mode used for short trips in urban areas, as it is often the case for airport access legs. To be able to derive this mode choice, a logit model has been calibrated. This logit model has provided good results when it comes to aggregated modal shares, both at an individual and at tariff zone level. However, the model is less accurate in the airport access legs, probably due to the influence of certain variables (e.g., value of time, travelling with luggage) that are different from the usual urban and metropolitan trips used to calibrate the model. This suggests the need for a tailored model for airport access, which provided a more accurate representation of air transport passengers' modal choices.
- Finally, to better characterise the passengers' trip purpose, the observed mobility patterns have been combined with machine learning techniques to classify trips into business and leisure. The proposed models are capable of predicting correctly the purpose of the majority of trips. This information will be of utmost importance to feed the long-distance transport simulation framework that will be developed in TRANSIT's WP5.

Regarding the characterisation of air transport supply, we have proposed a number of techniques to analyse airport connectivity and congestion. These techniques allow us to quantify to what extent an OD pair is well served and how resilient a trip is. This characterisation will be used in WP6 to analyse long-distance travel choices and passenger behaviour under disruption scenarios.

Appendix A: Data Factsheets

Orange Mobile Phone Data

General Information

Data source name: Orange Spain Mobile Phone Data

Last update of this file: 10/03/2021

Contact information: Alex Gregg (Nommon)

Support: Not applicable

Abstract

Mobile phone Call Detail Records contain the geolocated information of mobile phone users every time they make/receive a call, send/receive an SMS or use a data connection to access the Internet.

The key features of this data source are:

- capability to identify door-to-door trips
- big sample (more than 20% of the total population)
- passive collection, which allows the study of unpredicted events
- contains data of both roamers-in (non-Spanish people that connect to Orange Spain network) and roamers-out (Orange Spain clients travelling abroad)

Availability

Owner: Private (Orange)

Access conditions: Through private agreement

Data access: Batch

Data cost: Price specified in the data access agreement on a project-by-project basis

Access limitations: No limitations

Availability within the project: Only Nommon has access to the disaggregated data

Privacy/ Confidentiality issues: No issues, as data has been previously anonymised by Orange

Security issues: Data is considered as confidential information. Security requirements are specified in the data access agreement

State: Already Available

Link to data: Not applicable

Data Characteristics

Estimated size of the sample: 20-30% of Spanish total population (depends on the day)

Temporal scope: Open to the project needs

Geographical scope: Spain

Temporal granularity: Depends on the user usage of the mobile phone. Typically, one register every half an hour.

Geographical granularity: Mobile network antenna level (from 100-200m in urban areas to 1-2km in rural areas).

Delivery frequency: Hourly

Delivery Delay: One hour

Data format: CSV

Quality Issues

Geographical granularity of mobile phone data depends on the mobile network antenna density, giving spatial uncertainty of around 200m in big cities, but up to 2km in rural areas.

User profile information may not always be accurate, as the client (who is the person that appears in the Mobile Network Operator files) may not be the user of the mobile phone (teenagers who have their phone paid by their parents, for example).

When no network coverage maps are available, Voronoi areas are used as a proxy of antenna coverage, which is not the optimal approach (they do not take into account obstacles, or antenna technology).

Comments

As this data source also captures the mobility of the roamers that connect to the Orange Network in Spain, it will be explore if this can be exploited to determine mobility patterns and door-to-door trips across Europe.

Madrid Public Transport Smart Card Data

General Information

Data source name: Madrid public transport smart card data

Last update of this file: 10/03/2021

Contact information: Alex Gregg (Nommon)

Support: Not applicable

Abstract

This dataset contains all the registers of the Madrid public transport smart card users. A register is produced whenever a user enters the PT system, and, for some cases, when he gets out. This kind of information will be useful to calculate the mode of transport used to access/egress the airport.

Availability

Owner: Private (CRTM)

Access conditions: Through private agreement with CRTM

Data access: Batch

Data costs: Not applicable

Access limitations: Only for project purposes

Availability within the project: Only Nommon has access to disaggregated data

Privacy/ Confidentiality issues: No issues, as data has previously been anonymised by data provider

Security issues: Data is considered as confidential information. Security requirements are specified in the data access agreement

State: Already Available

Link to data: Not applicable

Data Characteristics

Estimated size of the sample: Around 60% of Madrid Public transport users

Temporal scope: Open to the project needs

Geographical scope: Madrid Region

Temporal granularity: Precise timestamp of the validation

Geographical granularity: GPS coordinates of the public transport stop

Delivery frequency: Not applicable

Delivery Delay: Not applicable

Data format: CSV

Quality Issues

For most of the public transport system, it is not needed for the user to use their smart card to get out of the system. Therefore, for many cases, it is only available the register that corresponds to the start of the trip. Also, up to now, only the locations of the intercity buses is available, and no location information about metro or train stops, which causes the dataset to not being completely useful.

EMMA surveys

General Information

Data source name: EMMA

Last update of this file: 10/03/2021

Contact information: Alex Gregg (Nommon)

Support: Not applicable

Abstract

EMMA is a survey carried out in the Spanish airport by AENA. The survey characterises the passenger characteristics and the purpose of the trip. The data is collected in periodic waves, through interception surveys, which provide relevant information about passengers: age, gender, reason for the trip, destination of the trip, frequency of the trip, place of residence, socioeconomic characteristics, etc. The Madrid Barajas ones, were carried out in two waves in June and another in November, and in total 18,699 passengers were surveyed. These surveys can help with the passenger characterisation and to characterise the trip purpose.

Availability

Owner: Private (AENA)

Access conditions: Through private agreement with AENA (project partner)

Data access: Batch

Data costs: Not applicable

Access limitations: Only for project purposes

Availability within the project: All TRANSIT project partners

Privacy/ Confidentiality issues: No confidentiality issues identified

Security issues: No security issues identified

State: Already Available

Link to data: Not applicable

Data Characteristics

Estimated size of the sample: Over 18,000 people were surveyed

Temporal scope: The surveys were carried out in two waves: June 2018 (from the 7th until the 13th) and November 2019 (from the 13th until the 20th)

Geographical scope: Madrid at a district level and the rest of the country at a municipal level

Temporal granularity: Twice a year

Geographical granularity: Madrid - Barajas Airport

Delivery frequency: Twice a year

Delivery Delay: Not applicable

Data format: CSV

Quality Issues

The survey only includes departures and transfers, which excludes the passengers arriving to the airport and not transferring. However, the trips are considered to be symmetric, i.e., the purpose of the trip is the same on the first trip and the return trip. Some survey fields may not be complete and may be subject to respondents' bias.

EDM 2018 survey

General Information

Data source name: EDM 2018 ("*Encuesta Domiciliaria de Movilidad*", i.e., Household Travel Survey)

Last update of this file: 10/03/2021

Contact information: Alex Gregg (Nommon)

Support: Not applicable

Abstract

The Mobility Household Survey is carried out by the Regional Transport Authority and analyses the Madrid residents' daily trips on a working day. It is based on 85,000 personal and telephone interviews carried out between February and May 2018. The data is considered to be useful for passenger and trip characterisation and to develop discrete mode choice models.

Availability

Owner: Public

Access conditions: Open source

Data access: Batch

Data costs: Free

Access limitations: No limitations identified

Availability within the project: All TRANSIT project partners

Privacy/ Confidentiality issues: No confidentiality issues identified

Security issues: No security issues identified

State: Already Available

Link to data: Not applicable

Data Characteristics

Estimated size of the sample: Over 85.000 people

Temporal scope: February to May 2018

Geographical scope: Madrid Region

Temporal granularity: Start time and end time for each trip

Geographical granularity: Madrid Transport Zoning

Delivery frequency: Mobility Household Survey, every 5 years

Delivery Delay: Not relevant (next EDM is planned for 2023, after the end of this project)

Data format: CSV

Quality Issues

Household surveys are produced every 5 to 10 years, with the last one being produced in 2018. This makes these data obsolete when used for some studies. Household surveys may be subject to respondents' bias. In addition, not all survey fields are fully complete.

OAG flight schedule

General Information

Data source name: OAG flight schedule

Last update of this file: 28/03/2021

Contact information: Daniel Delahaye (ENAC)

Support: Not applicable

Abstract

10 years of historical flight schedules at CDG airport. For each flight arriving or departing from CDG: scheduled departure time, scheduled arrival time, departure airport, arrival airport, carrier and aircraft type are known. These data are communicated by airlines and correspond to the offer provided to passengers.

Availability

Owner: OAG

Access conditions: Through private agreement

Data access: OAG flight schedule

Data costs: Free

Access limitations: No limitations identified

Availability within the project: Only ENAC has access to the disaggregated data

Privacy/ Confidentiality issues: No confidentiality issues identified

Security issues: Data considered confidential information

State: Already Available

Link to data: Not applicable

Data Characteristics

Estimated size of the sample: 4Mo

Temporal scope: 2 months of 2019

Geographical scope: Charles-de-Gaulle

Temporal granularity: Minutes

Geographical granularity: Not applicable

Delivery frequency: Weekly batch

Delivery Delay: Instant

Data format: CSV

RDPS data

General Information

Data source name: RDPS data

Last update of this file: historical days (no updates)

Contact information: Geoffrey Scozzaro (ENAC)

Support: Not applicable

Abstract

The data set is an extract of radar detection processed by the STR and sampled with a temporal granularity around the minute. It contains all the real departure and arrival flights at CDG airport within an operational day. For a departure flight, the origin airport, destination airport, callsign and real departure time are known. For an arrival flight, the origin airport, destination airport, callsign and real arrival time are known. The arrival time of flights at CDG is associated to the last trajectory 'point' provided by this system and the departure time corresponds to the first trajectory point provided by this system.

Availability

Owner: DGAC

Access conditions: Limited to DGAC services

Data access: Monthly batch

Data costs: Free

Access limitations: No limitations identified

Availability within the project: Only ENAC has access to the disaggregated data

Privacy/ Confidentiality issues: No confidentiality issues identified

Security issues: Data considered confidential information

State: Already Available

Link to data: Not applicable

Data Characteristics

Estimated size of the sample: 100ko

Temporal scope: 3 days of operation at Charles-de-Gaulle

Geographical scope: Charles-de-Gaulle

Temporal granularity: Minutes

Geographical granularity: Not applicable

Delivery frequency: Monthly batch

Delivery Delay: one month

Data format: CSV

Quality Issues

The estimated arrival time can be slightly in advance than the real arrival time due to radar detection and the temporal sampling. In the same way, the estimated departure time can be slightly later than the real departure.