Deliverable 3.3

Methodologies and Algorithms for Mobility Data Analysis

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815069.

Table of Contents

Table Index

Figure Index

Summary sheet

Deliverable No.	3.3
Project Acronym	MOMENTUM
Full Title	Modelling Emerging Solutions for Urban Mobility
Grant Agreement No.	815069
Responsible Author(s)	Antonio David Masegosa (Deusto) Jenny Fajardo (Deusto) Luis Enrique Díez (Deusto) Galder Martín (Deusto) Pablo Fernández (Deusto) Javier Burrieza (Nommon) Ignacio Martín (Nommon) Oliva García (Nommon) Georgia Aifandopoulou (CERTH) Josep María Salanova Grau (CERTH) Panagiotis Tzenos (CERTH)
Peer Review	Rodric Frederix (TML), Christian Heil (Regensburg), SanthanaKrishnan Narayanan (TUM), Stefan Sulzenbacher (REWAG)
Quality Assurance Committee Review	General Assembly
Date	16/12/2020
Status	Draft
Dissemination level	Public
Abstract	The general objective of this document is to describe the new data fusion and artificial intelligence algorithms developed for the extraction of the mobility patterns and travel choice behaviour drivers, and the new studies performed on emerging mobility solutions and the use of prominent big data sources. In this way, the main outcomes of this deliverable are two-fold. First, a set of new methodologies based on data fusion and Artificial Intelligence for the estimation of representative OD matrices, the comparison of OD matrices, the assessment of the potential demand that can be served by DRT or Ride-sharing systems, as well as the extraction of user trip diaries and the inference of home location and income

	level from shared-mobility operation data. Second, new studies concerning emerging mobility solutions focused on the adoption and use of shared mobility services in MOMENTUM case studies, and the utilization of big data sources to describe mobility patterns and to assess the effects of certain policies.
Version	Issue 2 Draft 1
Work Package No.	3
Work Package Title	Data Collection and Analysis
Programme	Horizon 2020
Coordinator	EMT Madrid
Website	www.h2020-momentum.eu
Starting date	01/05/2019
Number of months	36

Project partners

Organisation	Country	Abbreviation
EMPRESA MUNICIPAL DE TRANSPORTE DE MADRID SA	Spain	EMT
NOMMON SOLUTIONS AND TECHNOLOGIES SL	Spain	NOMMON
DIMOS THESSALONIKIS	Greece	THESS
ETHNIKO KENTRO EREVNAS KAI TECHNOLOGIKIS ANAPTYXIS	Greece	CERTH
STAD LEUVEN	Belgium	LEUVEN
TRANSPORT & MOBILITY LEUVEN NV	Belgium	TML
STADT REGENSBURG	Germany	REGENSBURG
TECHNISCHE UNIVERSITAET MUENCHEN	Germany	TUM
AIMSUN SL	Spain	AIMSUN SL
POLIS – PROMOTION OF OPERATIONAL LINKS WITH INTEGRATED SERVICES, ASSOCIATION INTERNATIONALE	Belgium	POLIS
UNION INTERNATIONALE DES TRANSPORTS PUBLICS	Belgium	UITP
UNIVERSIDAD DE LA IGLESIA DE DESUTO – ENTIDAD RELIGIOSA	Spain	UDEUSTO

Document history

Version	Date	Organisation	Main area of changes	Comments
Issue 1 Draft 1	30/11/2020	Deusto	Initial version	
Issue 2 Draft 1	16/12/2020	Deusto	All sections	Consortium internal review

List of acronyms

ICT	Information and Telecommunication Technologies
WP	Work Package
DRT	Demand Responsive Transport
AI	Artificial Intelligence
OD	Origin-Destination
KPI	Key Performance Indicator
SVM	Support Vector Machine
SVR	Support Vector Regressor
FCD	Floating Car Data
GDPR	General Data Protection Regulation
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
BT	Bluetooth
VRP	Vehicle Routing Problem
IDART	Integrated Dial a Ride problem
PDTSP	Pickup and Delivery Travelling Salesman Problem

Executive Summary

The overall goal of the MOMENTUM project is to develop a set of mobility data analysis and exploitation methods, transport models and planning and decision support tools able to capture the impact of emerging transport modes, such as shared mobility services, and ICT-driven behavioural changes on urban mobility, in order to support local authorities in the task of designing the right policy mix to exploit the full potential of emerging mobility solutions.

The general objective of this document is to **describe the new data fusion and artificial intelligence algorithms** developed for the extraction of the mobility patterns and travel choice behaviour drivers, and the **new studies performed on emerging mobility solutions and the use of prominent big data sources.** From this general objective, the specific goals posed for this document are:

- **Present the new data fusion and Artificial Intelligence methods and algorithms developed** and designed according to the test cases proposed in WP2 and to requirements from WP4 and WP5.
- **Validate the methodologies developed** in the four case studies established at MOMENTUM.
- **Perform new studies** related to **emerging mobility solutions** to characterise the adoption and use patterns of these services and to the **use of prominent big data sources** for describing mobility patterns.
- **Study the implications of the methodologies and analyses developed** for the four test cases defined in MOMENTUM.

The main outcomes of this document are the following:

- **A set of new methodologies based on data fusion and Artificial Intelligence for extracting and analysing mobility patterns**. This set of new methods were developed for:
    - Estimation of representative OD matrices
    - OD matrix comparison through a new structural similarity measure
    - Estimation of the potential demand that can be served by DRT or Ride-sharing systems.
    - Extraction of user trip diaries from shared-mobility operation data.
    - Home location and income level inference from shared-mobility operation data.
- **New studies concerning emerging mobility solutions and the use of prominent big data sources in the MOMENTUM case studies.** They focus on the adoption and use of shared mobility services, and the use of big data sources to describe mobility patterns and to assess the effects of certain policies. More specifically, the new studies are related to:
    - Comparison of OD matrices extracted from household survey data and mobile network data.
    - Shared-mobility role in restrictions on private vehicles to study the impact of emerging mobility solutions and also the technical opportunities and challenges that the use of big data sources entails for such assessments.
    - Shared-mobility adoption patterns.
    - Shared-mobility use.

1. Introduction

1.1. Scope and Objective

The overall goal of the MOMENTUM project is to develop a set of mobility data analysis and exploitation methods, transport models and planning and decision support tools able to capture the impact of emerging transport modes, such as shared mobility services, and ICT-driven behavioural changes on urban mobility, in order to support local authorities in the task of designing the right policy mix to exploit the full potential of emerging mobility solutions.

In order to achieve this general goal, one of the particular objectives set was to characterise emerging activity-travel patterns, by profiting from the increasing availability of high-resolution spatio-temporal data collected from personal mobile devices and digital sensors. This document falls under this particular objective, that is also the main objective of the WP3 "Data Collection and Analysis". The final and most important part of this WP aims at developing new data fusion and artificial intelligence algorithms for the extraction of the mobility patterns and travel choice behaviour drivers of different population groups. **The general objective of this document is to describe the new data fusion and artificial intelligence algorithms developed as well as the experimentation and data analysis carried out to validate their correct performance**. In this way, the specific objectives posed for this document are:

- **Present the new data fusion and Artificial Intelligence methods and algorithms developed.** The design of these methodologies has been made taking into account the test cases proposed in task T2.3 and collected in deliverable D2.2 "Specification of the MOMENTUM Test Cases", and also the requirements from WP4 and WP5.
- **To validate the methodologies developed in the four case studies established at MOMENTUM**, that is, Madrid, Leuven, Regensburg and Thessaloniki. This was done using the data sources collected in Task 3.1 of WP3 and documented in deliverable D3.1 "Data Inventory and Quality Assessment".
- **Perform data analyses** related, on the one hand, to **emerging mobility solutions** (e.g. car-sharing, bike-sharing) to **characterize the adoption and use patterns of these services** in order to better understand the profile of the people who consume these services and the type of trips they are used for; and on the other hand, to the **use of prominent big data sources for describing mobility patterns**. This information is very useful for WP4, as it will support the modelling of these transport modes, especially in terms of demand.
- **To study the implications that the methodologies developed and the results obtained** have for the test cases that were defined for each of the four cities involved in MOMENTUM.

1.2. Structure of the document

The next part of the document is structured in four sections. **Section 2 describes the new methodologies developed** following the specifications of the MOMENTUM test cases and the requirements from WP4 and WP5. The content of this section is as follows:

- **Section 2.1** presents a methodology that aims to **estimate a representative OD matrix by the fusion of a set of similar OD matrices** using statistical learning techniques.
- **Section 2.2** shows a **new measure for comparing OD matrices**, concretely a new **structural similarity measure** (i.e. it takes into account the underlying patterns and correlations among OD pairs).
- **Section 2.3** contains the third methodology developed. This is a **framework to estimate the potential demand that can be served by a DRT or Ride-sharing system**.

- **Section 2.4** presents a **generic approach to loading, homogenising, and modelling data from shared mobility operators**.
- **Section 2.5** describes **new methodologies developed for the profiling of shared mobility users, including home location and user income**.

Section 3 is dedicated to the validation of the methodologies listed above in the four MOMENTUM case studies and also to **describe the data analyses related to the adoption and use of emerging mobility solutions**. As shown below, the content has been divided into four subsections, one for each case study:

- **Section 3.1** contains the **validations and results related to the Madrid Case Study**. This includes a comparison between surveys and mobile network data to obtain OD matrices, an analysis of the adoption and use of shared mobility services in the city, and the demonstration of the techniques for estimating representative OD matrices and comparing them using those obtained from mobile network data.
- **Section 3.2** is dedicated to the **Leuven Case Study**. This includes the analysis of survey data in relation to car-sharing adoption patterns and the willingness to use it, making use of machine learning techniques. The relation with car-ownership is explored more in detail.
- **Section 3.3** presents the validations and analyses carried out in the **Regensburg Case Study**. The section presents the studies on the adoption and use patterns of the car-sharing system in the city, exploiting survey and operation data
- **Section 3.4** is dedicated to the last use case, **Thessaloniki**. It describes the validation of the new method for estimating representative OD matrices using data from GPS traces of taxis, the validation of proposed methodologies for the estimation of the potential demand to be served by a DRT or ride-sharing service, and a second validation of the structural similarity measure for comparing OD matrices using OD matrices extracted from floating car data from taxis.

Section 4 aims at **analysing the implications that the methodologies developed, and results obtained have for the four MOMENTUM test cases** described in deliverable D2.2. Finally, **Section 5 discusses the main conclusions drawn from this document** and its relation to the following work packages, concretely, WP4 and WP5.

1.3. Reference and applicable documents

Applicable documents:

[I] Grant Agreement No 815069 MOMENTUM – Annex 1 Description of the Action.

[II] MOMENTUM Consortium Agreement, Issue 1, April 2019.

[III] MOMENTUM D1.1 Project Plan, June 2019

[IV] MOMENTUM D1.2 Data Management Plan and Open Data Policy, November 2019

[V] MOMENTUM D2.2 Specification of MOMENTUM Test Cases, February 2020

[VI] MOMENTUM D3.1 Data Inventory and Quality Assessment, March 2020

[VII] MOMENTUM D3.2 MOMENTUM Data Repository, June 2020

Reference documents:

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Zheng, X. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. Retrieved from http://arxiv.org/abs/1603.04467

Aguilera-García, Á., Gomez, J., & Sobrino, N. (2020). Exploring the adoption of moped scooter-sharing systems in Spanish urban areas. Cities, 96, 102424.

Arias-Molinares, D., & Palomares-García, J. C. (2020). Shared mobility development as key for prompting Mobility as a Service (MaaS) in urban areas: the case of Madrid. Case Studies on Transport Policy, 8(3).

Ashok, K., & Ben-Akiva, M. E. (2002). Estimation and prediction of time-dependent origin-destination flows with a stochastic mapping to path flows and link flows. Transportation Science, 36(2), 184–198. https://doi.org/10.1287/trsc.36.2.184.563

Bassolas, A., Ramasco, J. J., Herranz, R., & Cantú-Ros, O. G. (2019). Mobile phone records to feed activity-based travel demand models: MATSim for studying a cordon toll policy in Barcelona. Transportation Research Part A: Policy and Practice, 121, 56-74.

Becker, H., Balac, M., Ciari, F., & Axhausen, K. W. (2020). Assessing the welfare impacts of Shared Mobility and Mobility as a Service (MaaS). Transportation Research Part A: Policy and Practice, 131, 228-243.

Becker, H., Ciari, F., & Axhausen, K. W. (2017). Comparing car-sharing schemes in Switzerland: User groups and usage patterns. Transportation Research Part A: Policy and Practice, 97, 17-29. doi: 10.1016/j.tra.2017.01.004

Behara, K. N. S., Bhaskar, A., & Chung, E. (2020a). A novel approach for the structural comparison of origin-destination matrices: Levenshtein distance. Transportation Research Part C: Emerging Technologies, 111, 513–530. https://doi.org/10.1016/j.trc.2020.01.005

Behara, K. N. S., Bhaskar, A., & Chung, E. (2020b). Geographical window based structural similarity index for origin-destination matrices comparison. Journal of Intelligent Transportation Systems: Technology, Planning, and Operations, 1–22. https://doi.org/10.1080/15472450.2020.1795651

Behara, K., Bhaskar, A., & Chung, E. (2018). Classification of typical Bluetooth OD matrices based on structural similarity of travel patterns: Case study on Brisbane city. Proceedings of the Annual Meeting The Transportation Research Board (TRB) 97th Annual Meeting.

Bricka, S., & Bhat, C. R. (2006). Comparative analysis of Global Positioning System–based and travel survey–based data. Transportation Research Record, 1972(1), 9-20.

Ceccato, R. (2020). Switching intentions towards car sharing.

Clewlow, R. R. (2016). car-sharing and sustainable travel behavior: Results from the San Francisco Bay Area. Transport Policy, 51, 158-164. doi:10.1016/j.tranpol.2016.01.013

Cools, M., Moons, E., & Wets, G. (2010). Assessing the Quality of Origin–Destination Matrices Derived from Activity Travel Surveys. Transportation Research Record: Journal of the Transportation Research Board, 2183(1), 49–59. https://doi.org/10.3141/2183-06

CRTM (2019). Encuesta domiciliaria de Madrid [Spanish]. https://www.crtm.es/conocenos/planificacion-estudios-y-proyectos/encuesta-domiciliaria/edm2018.aspx

Djukic, T., Hoogendoorn, S., & Van Lint, H. (2013). Reliability Assessment of Dynamic OD Estimation Methods Based on Structural Similarity Index.

DS Chen, R. B. (2010). Applied integer programming. Wiley Online Library.

E. Martin, S. Shaheen, J. Lidicker. Impact of carsharing on household vehicle holdings: Results from north american shared-use vehicle survey. Transport. Res. Rec.: J. Transport. Res. Board, 2143 (2010), pp. 150-158

E. Martin, S. Shaheen. The impact of carsharing on public transit and non-motorized travel: an exploration of north american carsharing survey data Energies, 4 (11) (2011), pp. 2094-2114

Enoch, M., Potter, S., & Parkhurst, G. a. (2006). Why do demand responsive transport systems fail? Transportation Research Board 85th Annual Meeting.

Fan Zhou, Zuduo Zheng, Jake Whitehead, Robert K. Perrons, Simon Washington, Lionel Page, Examining the impact of car-sharing on private vehicle ownership, Transportation Research Part A: Policy and Practice, Volume 138, 2020, Pages 322-341,

Fishman, E., Washington, S., & Haworth, N. (2014). Bike share’s impact on car use: Evidence from the United States, Great Britain, and Australia. Transportation Research Part D: Transport and Environment, 31, 13-20.

H. Becker, F. Ciari, K.W. Axhausen, Measuring the car ownership impact of free-floating car-sharing - A case study in Basel, Switzerland, Transp. Res. Part D: Transp. Environ., 65 (2018), pp. 51-62

Häll, C. A. (2009). The Integrated Dial-a-Ride Problem. Public Transp 1.

Jakkula, V. (2006). Tutorial on support vector machine (svm).

Jakobsson Bergstad, C., Ramos, E., Chicco, A., Diana, M., Beccaria, S., Melis, M., Rodenbach, J., Matthijs, J., Nehrke, G., Loose, W., 2018. STARS Shared mobility opporTunities And challenges foR European citieS: Deliverable D4.1 - The influence of socioeconomic factors in the diffusion of car sharing 243.

Jiao, J., Bischak, C., & Hyden, S. (2020). The impact of shared mobility on trip generation behavior in the US: Findings from the 2017 National Household Travel Survey. Travel Behaviour and Society, 19, 1-7.

Jin, C., Nara, A., Yang, J., & Tsou, M. (2020). Similarity measurement on human mobility data with spatially weighted structural similarity index (SpSSIM). Transactions in GIS, 24(1), 104–122. https://doi.org/10.1111/tgis.12590

Joshua B. Tenenbaum, V. d. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science.

Kabra, A., Belavina, E., & Girotra, K. (2020). Bike-share systems: Accessibility and availability. Management Science, 66(9), 3803-3824.

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. ACM Computing Surveys (CSUR), 52(4), 1-36.

Kim, S.-J., Kim, W., & Rilett, L. R. (2005). Calibration of Microsimulation Models Using Nonparametric Statistical Techniques. Transportation Research Record: Journal of the Transportation Research Board, 1935(1), 111–119. https://doi.org/10.1177/0361198105193500113

Kortum, K., and R. Machemehl. Free Floating Carsharing Systems: Innovations in Membership Prediction, Mode Share, and Vehicle Allocation Optimization Methodologies. Report SWUTC/12/476660-00079-1. Southwest Region University Transportation Center, Center for Transportation Research, University of Texas at Austin, 2012

Lee, Wenke, and Dong Xiang. "Information-theoretic measures for anomaly detection." Proceedings 2001 IEEE Symposium on Security and Privacy. S&P 2001. IEEE, 2000.

Liao, F., Molin, E., Timmermans, H., & van Wee, B. (2020). Carsharing: The impact of system characteristics on its potential to replace private car trips and reduce car ownership. Transportation, 47(2), 935-970.

Luxburg, U. V. (2007). A tutorial on spectral clustering. Statistics and computing.

M. Belkin, P. N. (2003). Laplacian Eigenmaps for Dimensionality Reduction and Data Representation.

M.E. Bruni, F. G. (2014). Designing robust routes for demand-responsive transport. Transportation Research Part E.

Ma, X., Yuan, Y., van Oort, N., & Hoogendoorn, S. P. (2020). Bike-sharing systems' impact on modal shift: A case study in Delft, the Netherlands. Journal of Cleaner Production, 259.

Martin, E., Shaheen, S., 2016. Impacts of car2go on Vehicle Ownership, Modal Shift, Vehicle Miles Travelled, and Greenhouse Gas Emissions: An Analysis of Five North American Cities. Berkeley.

Meropi, Pavlidou, Christoforos Bikos, and Zioutas George. "Outlier detection in skewed data." Simulation Modelling Practice and Theory 87 (2018): 191-209

Patrick Jochem, Dominik Frankenhauser, Lukas Ewald, Axel Ensslen, Hansjörg Fromm, Does free-floating carsharing reduce private vehicle ownership? The case of SHARE NOW in European cities, Transportation Research Part A: Policy and Practice, Volume 141, 2020, Pages 373-395,

Primerano, F., Taylor, M. A., Pitaksringkarn, L., & Tisato, P. (2008). Defining and understanding trip chaining behaviour. Transportation, 35(1), 55-72.

R. Katzev, Car sharing: A new approach to urban transportation problems. Anal. Soc. Issues Public Policy, 3 (1) (2003), pp. 65-86

Rémy Chevriera, ∗. A. (2011). Solving a dial-a-ride problem with a hybrid evolutionary multi-objective. Applied Soft Computing.

Ros-Roca, X., Montero, L., Schneck, A., & Barceló, J. (2018). Investigating the Performance of SPSA in Simulation-Optimization Approaches to Transportation Problems. In Transportation Research Procedia (Vol. 34, pp. 83–90). Elsevier B.V. https://doi.org/10.1016/j.trpro.2018.11.017

Ruiz de Villa, A., Casas, J., & Breen, M. (2014). OD matrix structural similarity: Wasserstein metric.

S. Le Vine, J. Polak. The impact of free-floating carsharing on car ownership. Early-stage findings from London, Transport Policy, 75 (2019), pp. 119-127,

Schmöller, S., Weikl, S., Müller, J., & Bogenberger, K. (2015). Empirical analysis of free-floating carsharing usage: The Munich and Berlin case. Transportation Research Part C: Emerging Technologies, 56, 34-51.

Shen, Y., Zhang, X., & Zhao, J. (2018). Understanding the usage of dockless bike sharing in Singapore. International Journal of Sustainable Transportation, 12(9), 686-700.

Shlens, J. (2014). A Tutorial on Principal Component Analysis.

Tavassoli, A., Alsger, A., Hickman, M., & Mesbah, M. (2016). How close the models are to the reality? Comparison of transit origin-destination estimates with automatic fare collection data.

Tourassi, Georgia D., et al. "Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms." Medical Physics 34.1 (2007): 140-150.

Van der Maaten, L., & Hinton, G. (2008). Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research.

Van Vuren, T., & Day-Pollard, T. (2015). 256 shades of grey-comparing OD matrices using image quality assessment techniques. In 2015 Scottish Transport Applications Research Conference Proceedings.

WERABHAT MUNGTHANYA1, S. P. (2019). Constructing Time-Dependent Origin-Destination

Yoon, T., Cherry, C. R., & Jones, L. R. (2017). One-way and round-trip car-sharing: A stated preference experiment in Beijing. Transportation Research Part D: Transport and Environment, 53, 102-114. doi: 10.1016/j.trd.2017.04.009

Zack, R. (2019). A Data Standard for New Mobility. Institute of Transportation Engineers. ITE Journal, 89(6), 26-28.

2. Description of the new methodologies developed

2.1. OD representative matrix estimation from geolocated data

2.1.1. Motivation and objectives

The proliferation of mobile devices and sensors has provided the transport sector with an opportunity to collect high quality geolocated data about people and vehicle movements. One of the most important mobility data sources is the recording of trajectories by different means, ranging from tracking vehicles at static locations (e.g. detecting Bluetooth devices) or tracing them along the whole route (e.g. using a telematics system). Both approaches are also applicable for individuals (e.g. tracking pedestrians through beacons or tracing travellers through GPS data from mobile apps):

- The data obtained in the first case (tracking) has the advantage that the only infrastructure needed is a set of detectors (Bluetooth devices detectors) in static locations, but it needs a complex data processing in order to select the correct samples (for instance, if the study focuses on vehicles, the sample will also detect pedestrians and PuT passengers) and most important, it does not detect origin and destination locations.
- The data obtained in the second case (tracing) has the advantage that it detects with accuracy origin and destination, but the need for equipment is larger in the case of vehicle tracing, given that all the fleet of vehicles should be equipped with data loggers (logging Floating Car Data - FCD).

In both cases, there are two important aspects that have to be considered when using the data:

- The sample from the population that will be analysed depends on the penetration of technology and the drivers' or travellers' behaviour. In the tracking case for vehicles, there is a need to have a Bluetooth device in the car, to be open and visible, which in the new cars it mostly applies. In the tracing case for vehicles, there is a need to have a GPS data logger and to be used by the driver, which in professional fleets it usually happens. In the case of individuals, the sample size and potential biases will depend on the population groups using the technology generating the data.
- There are privacy concerns with regards to personal information being recorded (GDPR), which can be solved by anonymising the identities or the MAC ids. Additional anonymization can be added to the tracing case in order to avoid back-tracking of travellers. For instance, the information about the detailed origin and destination locations can be diffused by not logging the first and last 100 meters.

The extraction of travel demand information from these sources requires the development of methodologies able to process, transform and interpret the data. A very common initial goal for these analyses is to obtain an OD matrix from a given area and period of time. Each cell in the matrix represents the number of trips performed from an origin zone to an origin-destination. The obtention of such matrices involves the accurate identification of movements from the registers, spatial and temporal aggregations as well as sampling scaling techniques. The complexity of these steps varies from one data source to another, depending on the characteristics of the raw data.

One of the aspects that have to be taken into account when analysing OD matrices is that part of the OD pairs that show no trips ('blank cells') are actually wrong, since some trips would have occurred during the study period, but they were not included in the sample obtained through the applied data collection. Even if this effect is significantly alleviated thanks to the use of larger samples from big data sources than survey data, it is consubstantial to the sampling approach. Transport modellers try to overcome this issue by building representative OD matrices for a certain type of day (e.g. 'average working day'). If several days are supposed to show very similar mobility patterns, it is reasonable to aggregate the OD matrices resulting from the observation of these days to gain variability in the results. This is usually approached through standard average, but this does not directly address the effect of zero cells in the OD matrices. There are alternative methods that take into account asymmetries in distributions that may be useful for this purpose.

In this section, a methodology for estimating a representative OD matrix from geolocated data is provided, tailored to the features observed in mobility OD matrices related to asymmetrical distributions. The methodology has been oriented towards the analysis of trajectories of a taxi fleet, but it can also be applied to other vehicle or individual mobility databases. The methodology will be applied to daily and peak hour matrices from the GPS data generated by a taxi fleet in Thessaloniki, and to daily and peak hour matrices from the mobile network data describing overall mobility of travellers in Madrid.

2.1.2. State of the art

In general, OD matrices capture the citizens' mobility behaviour of a specific area in a specific time period and thus, different time periods (e.g. hourly, daily ODs) reflect different instances of the mobility situation. There is no appropriate or inappropriate time basis for an OD matrix, but each variant provides a pattern that responds to a different aspect of mobility. OD matrices play a central role in transport simulations, and they are used as the main tool for the evaluation of citizens' urban behaviour.

Furthermore, apart from their contribution to the identification of mobility patterns, OD matrices can also be used for transport analyses in different ways. As they represent mobility patterns, it would be interesting to extract knowledge about the changes in their structure over time. More specifically, it would be useful to take multiple instances of OD matrices across time (e.g. hourly time step) and recreate a representation of matrix/matrices that compress as much as possible the prior information. In parallel, these new OD matrices should augment the differences in mobility situations as much as possible. A typical OD matrix for a given time step is an O by D matrix, where O is the number of origin regions and D the number of destinations. So, each OD matrix instance of each time step can be treated as an observation. In this way, the problem takes a form similar to representation learning, dimensionality reduction, and clustering problems.

The first step for the evaluation of the relation between the OD matrices and the above-mentioned problems is to define a similarity measure. Given that similarity, we can derive the groups that correspond to different mobility modes. So far, most approaches consider the case of measuring dissimilarities among matrices. These dissimilarities are produced from common measures like mean squared error (MSE), root mean squared error (RMSE) or mean absolute error (MAE). The transportation literature presents an extensive and deep study of this topic, and more similarity measures like Mean Structural SIMilarity index (MSSIM), Wasserstein metric, or Levenshtein distance measure are also identified. A more detailed review about similarity measures for OD Matrix comparison is given in Section 2.3, where a new methodology of this type that has been developed within this project is presented. The interested reader is also referred to (KNS Behara, 2020) for an extensive analysis and literature review about the similarity measures mentioned above.

It is important to explain how this work can be valuable for information extraction about typical urban transport situations based on OD estimations. A typical point of view derives from clustering approaches; the basic statistical features of OD groups can be explored and compared to other groups. The more augmented the differences among clusters are, the greater the success of the clustering process. The similarity measure mentioned above is an element of high importance in a good clustering representation. To be more accurate, these similarities can construct an N by N similarity matrix, where N is the number of time steps, and each OD pair has its own similarity. This similarity is used as a distance by the different clustering algorithms. For example, k-means can use these similarities instead of computing Euclidean distance. One more interesting approach is the method of Spectral

Xlustering (Luxburg, 2007). In Spectral Clustering, the similarity matrix that fulfils symmetric and non-negativity requirements can directly be used as the affinity matrix computed by this method.

In all the aforementioned methods, we took OD instances and grouped them. However, there is one more approach worth mention that derives from the search in a different dimension of these stacked OD matrices. This approach uses the time dimension Z and reduces it on lower dimension space 3. So, if Z dimension refers to hours, the new dimensions can compress this information in 3-dimensional spaces. Each coordinate contains a piece of the total time spectrum, and its values have the new representation of the matrix. The most widely known algorithm for this dimensionality reduction is the principal component analysis (PCA) that projects data in a lower dimension by keeping as much uncertainty as possible in the new dimension space (Shlens, 2014). This field is an extensively studied area in statistical learning called manifold learning. It has a lot of powerful algorithms, such as the non-linear extension of kernel-PCA and generalizations like Isomap algorithm (Joshua B. Tenenbaum, 2000). Other important methods are the Spectral Embedding method which is closely related to Spectral Clustering method (M. Belkin, 2003) and the t-SNE algorithm (van der Maaten & Hinton, 2008).

The method developed in the MOMENTUM project tries to explore a new way of aggregating stacked OD matrices with direct statistical learning methods. In fact, it uses the Support Vector Machine (SVM) (Jakkula, 2006) formulation in order to give a good estimate of the representative OD matrix. Our method aims to make a direct projection of the 3D matrix (origin, destination, time period) generated by the set of OD matrices calculated at the same time interval for various days to a 2D matrix (origin, destination) with the use of a variant of SVM. The importance of the method stands on the sophisticated and well-designed optimization variant of primal SVM problem and the generality of the solution.

2.1.3. OD matrix estimation for a concrete time interval

FCD from taxi trips contains various series of locations (latitude and longitude) and temporal references (timestamp). Additional data such as speed, orientation, driver id, vehicle id, customer flag, stand flag… can be included depending on the telematics system.

The first step is to build the individual trajectories from the raw FCD. For this, the customer flag, which declares the presence of a traveller inside the taxi is used to split all trips. For the trips in which there is a customer inside, the origin and destination locations are stored together with the timestamps at both locations.

The second step is to map-match all the origins and destinations to the reference zoning system. Thus, transforming origin and destination coordinates into origin and destination zones.

In the third step, a set of hourly OD matrices are created by counting the number of trips between each OD pair within the respective time interval (the timestamp of the trip start is used).

2.1.4. OD matrix combination

The matrices for the time intervals that present a high similarity are combined into a representative matrix by integrating the information coming from all the values of each OD pair cell. The focus here is which metric should be used. Two methods for estimating one representative value out of the set of values of each OD pair cell are compared. In order to estimate a representative traffic flow value for the element OD (i,j) of the final OD matrix, we used the classic mean of the respective elements of the daily ODs that we computed from the data as well as a new estimate called ε-med which is suggested as a local estimate for data that are not necessarily Gaussian distributed, contrary to the mean value.

2.1.4.1. Mean value as estimator

The first method computes the mean value of each OD pair cell and accepts it as a measure of the element OD (i, j) of our final OD matrix. The mean value is estimated through Maximum Likelihood estimation, assuming that the distribution of the values is Gaussian. However, there is no guarantee that the data will even be symmetrically distributed. In practice, the resulting distribution will probably at least be asymmetrical and exhibit a level of skewness, mainly because of the high number of zeros usually present in OD matrices. In some preliminary experiments, we assess this skewness for OD matrices in Thessaloniki, and we observed that 95% of the cell were zeros, thus creating largely asymmetrical distributions.

2.1.4.2. ε-med SVM as estimator

Consequently, in order to estimate a representative local estimate for the final OD cells, a more inclusive averaging method that takes into account the asymmetries of the data, if present, and includes them in the estimation process if required. Thus, a new local estimate that has been recently proposed (Meropi, 2018) for both skewed and non-skewed distributed values, the ε -median, was used.

The ε-med is a Support Vector Regression (SVR) Mathematical Programming optimization problem whose solution is, among others, a new local estimate aiming at replacing the mean value. This local estimate does not presuppose that the data follow the Gaussian distribution, and therefore, it is more suitable for asymmetrical distributions (Figure 1). This new estimate is essentially an ε-insensitive Support Vector Machine (SVM) estimator. Thus, it does not resort to over-fitting on the data. In order to estimate an optimal value for parameter ε, a new constraint is added, resulting in the parameter ε being converted into a decision variable. Therefore, its value is estimated through the solution of the optimization problem and is not required beforehand. The new model is the following:

$$
\begin{gathered}
\text{minimize } \Sigma_i^n \xi_i \\
\text{subject to:} \\
\varepsilon\text{-med} - y_i \leq \xi_i + \varepsilon \\
-\varepsilon\text{-med} + y_i \leq \xi_i + \varepsilon \\
\xi_i \geq \varepsilon
\end{gathered}
$$

This method uses the new local estimate ε-med with the intent to include the data in the multivariate area of margin, where ξ_i is a measure of error outside this margin. As a result, observations that display errors smaller than or equal to ε contribute to the cost function. A trade-off relationship has been established between these observations and the rest ones that display errors larger than ε and are reduced by ε. In other words, all data points will contribute to the SVM cost function by ε. In an optimum solution, the value of the error will act as a trade-off between the larger deviations and the goal to rely more on data from the skewness direction. However, since this is a statistical method and the non-zero values are limited due to the sparsity of the matrices, the ε-med was first applied to the non-zero values, and then the mean value of the ε-med and the rest zero-elements were estimated. Thus, more representative values for each element of the final OD matrix are expected through this Support Vector Machine Optimization method, which is not only optimal for normally distributed values, but it is also suitable for skewed distributions and limited data, such as in the case study of Thessaloniki.

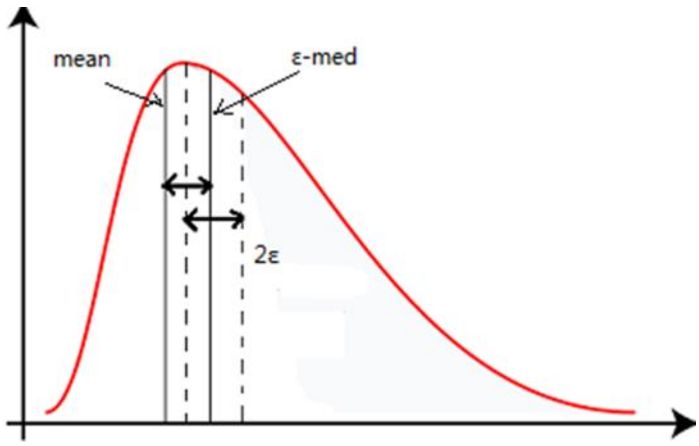


Figure 1 . Mean and ε-med estimators in a skewed distribution

The estimated interval |ε-med - ε| as the solution to the problem, is expected to be transferred from the high density area, where the deviations are the smallest, towards the skewness direction, offering a more representative local estimate for the data as it is relying (Figure 1), more than the traditional mean value, on a shifted portion of the data, towards the direction of skewness. In our case, this can be translated as the local estimate relying less on the high-density data, which obviously are the numerous zero values, and more on the sparse non-zero values which are low in frequency.

2.1.4.3. Comparison of the two methods

Since there is no ground truth available to compare the OD matrices obtained using the different metrics, the comparison should be made between the obtained matrix and the set of matrices used for their estimation. It is assumed that the final representative OD should be closer to the initial ODs. The metric selected for comparing the two matrices is the conditional entropy (Lee, 2000) since the real ODs are available, and we want to compare the information provided by them to the information provided by the representative matrix as the final representative OD will supposedly be suitable to substitute the real ODs and be close to them. In Information Theory, conditional entropy H(Y/X) describes the amount of information needed to describe one random variable Y, or the uncertainty referring to this variable Y, while the value of another random variable X is known.

$$H(Y/X) = \Sigma_{x,y}\, p(x,y)\, \log(p(x,y)/p(x))$$

Conditional Entropy measures the information content of the outcome Y, assuming that X is already known. The conditional Entropy H(Y/X) = 0 if and only if the value of Y is completely determined by the value of X. It can be stated that smaller entropy values mean more relation between the OD matrices and the representative one.

Conditional Entropy has been utilized as a measure of proximity and similarity before. In (Tourassi, 2007), conditional entropy was one of the similarity measures compared for the purposes of content-based retrieval and detection of masses in screening mammograms, in the category of Information-theoretic similarity measures. In (Lee, 2000), Conditional Entropy was used to recognize non-matching data that suggested for anomalies and possible intrusion in Computer and Network Science. Finally, in (Zhuang, 2015), conditional entropy was the basic metric applied for the selection and ranking of heart compartments and tissues.

In addition to entropy, the following metrics are used to compare the resulting matrices.

- Total number of trips
- Average number of trips per OD pair
- Total number of zeros
- Distribution of trip values

2.2. Trips clustering for on-demand and sharing mobility systems

2.2.1. Motivation and objectives

Managing transportation operations is a challenge for transportation managers and policymakers. Without metrics controlling the performance of operations in mobility services, evaluation of the interventions would not be possible. Thus, KPIs (Key Performance Indicators) help to get insights in evaluating transportation services efficiency and feasibility. Due to the lack of extensive literature and research progress on this topic, the presented methodology creates a complete framework that could support transport planning analysis and decision making for implementing DRT in a city.

Demand Responsive Transport (DRT) is described as a service that aims to serve passengers trips taking into account their commuting habits, providing flexible and safe trips. Therefore, it is important to distinguish which districts can benefit more from these services.

The main goal of this framework is to provide a set of tools that support the planner’s decision for appropriate transport infrastructure design and operation. More specifically, the presented method analyses some of the main features of the area, both operational and topological, to provide a clearer representation of opportunities and difficulties on the development of DRT. The first step is to set the KPIs, which can provide an insight of their feasibility in the examined area.

2.2.2. KPIs to assess the suitability of the implementation of an on-demand system

The calculation of the KPIs is a complicated procedure due to the combination of information coming from different sources. Some of those KPIs are related to demographical or topological issues while others are related to socioeconomic, public infrastructure or operational attributes of each area. Below, the parameters used for developing the KPIs are described.

- Socio-demographic characteristics (traveller): Age, Income, Car ownership, Driving license, Proximity on public or shared transport services.
- Region characteristics: Land use, Industrial areas, rent price, Shopping and lifestyle areas (malls, shopping streets, parks and bars), Business regions, Universities or areas generally preferred by students.
- Public transport services in the area: Parking space availability and/or price, Type of region (urban, suburban, periurban, exurb)
- PuT/sharing services operational characteristics: Average waiting time, in-vehicle time and access time, PuT headway, Sharing Services availability, Travel time reliability.

The parameters mentioned above are coming from different data sources, such as trip trajectory data (FCD, BT, mobile data etc.), network topology and stops and timetables of bus lines.

2.2.3. Formulations for the KPIs for a DRT system

In order to evaluate the feasibility of the implementation of a DRT mobility system, it is important to define indicators that can monitor the level of satisfaction of users with the new services. The identification of the indicators is based on the assessment of the total time users spend travelling, which is composed of the following:

- Access Time as a function of the number of service points, road network characteristics and operating system capabilities, topology of the area
- Waiting Time as a function of the inter-arrival time, size of the fleet
- In-vehicle Time as a function of the type of vehicle, congestion, number of service points (bus stops), topology of the area
- Uncertainty Factors as a function of the congestion, mechanic failure, accidents
- Monetary cost as a function of the minutes a passenger has to work to pay this route

The aim of a DRT system is to decrease the total travel time of passengers. Thus, the primal goal is to detect which of elements of Total Time function listed above can be improved.

KPI name	Stands For	Comments
Nts	Number of transport service	The different modes each area provide to passenger. (e.g. car, bike, bus, metro, etc.)
Nbl	Number of bus lines	The number of lines serving that area. (e.g. Area 2 by 3, 4, 34, 2 bus lines)
Nbs/ km^2	Number of bus lines / km ^2	It describes the density of stops.
IT	Interarrival time	Small IT means less waiting time and small IT standard deviation leads to a more reliable system.
Npa	Number of passengers arrivals / (km^2*hours)	More sparse demand usually combined with poor public transport infrastructure so that region is more suitable for DRT services.
Npla	number of passengers leave the area / (km^2*hours)	Has a similar behaviour as Npa
LU	Land Use	(Industrial, shopping, entertainment)

Table 1. KPIs for a DRT system

2.2.4. Demand served by DRT services

During the last decades, a lot of works have tried to develop efficient solutions, mostly for the operational part of Demand responsive systems (DRT). Therefore, literature presents many efficient solutions and useful variants of DRT problem. Most of those solutions stand as special instances of more general problems like Vehicle Route Problem (VRP) and it's a sub-category of Dial-a-Ride problem (DARP). Unfortunately, most DRT systems have failed after being implemented. There are many reasons explain that failure, (Enoch, Potter, & Parkhurst, 2006) but crucial role seems to play the wrong initial planning of the system. Other factors that lead to DRT failure derive either from an inability to cooperate with existing transport providers or the inefficient consideration of public authorities' restrictions. All those elements compose an uncertain and complicated environment for the implementation of a DRT service.

However, there are some studies that try to embed those uncertainties in operational optimization models (M.E. Bruni, 2014), (Rémy Chevriera, 2011) considering multiple aspects of DRT costs or the robustness of the solution. Other studies also try to consider the integration of DRT lines with pre-existing infrastructure (Häll, 2009) to avoid some of the public authorities' restrictions. Unfortunately, the crucial factor of the expected demand and the uncertainty it produces is not considered in the literature. In addition, all those studies end up with very complex and inefficient solutions. In contrast, that work achieves to integrate both operational and demand parts in a bigger framework that allows higher interpretability on different aspects of a DRT system. The role of demand estimation is critical for the successful implementation of such services, much more sensitive to it than conventional modes, and good planning requires an unbiased and accurate demand estimation. Current work takes place with FCD from taxi trips, but as more data sources for demand are available, the more precise and reliable is the planning of the DRT line.

Once districts suitable for DRT services are identified, the next step of the procedure is to define the framework that could help to plan and define the operational schedule of a DRT line and therefore the demand that will serve. The method applied in the case study, using taxi trips as a proxy of disaggregated demand, creates the optimization problem for planning the DRT service. The main idea is to create a simple procedure with defined steps. Figure 2 gives a general description of that process flow.

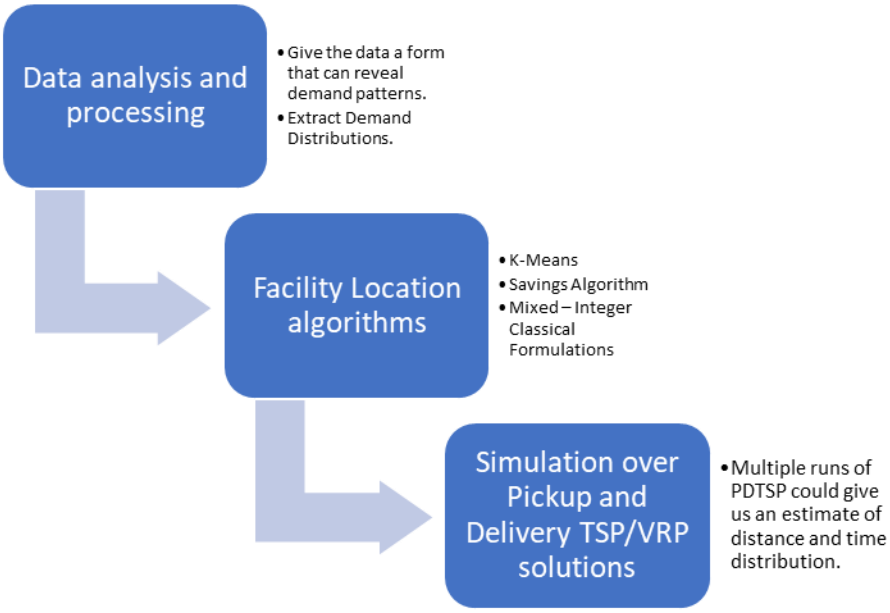


Figure 2. DRT optimization process flow

The first step is the pre-processing of data and appropriate methods that could extract probability distributions for each of the DRT stops. The next step is to define the stops that the DRT line would serve by solving the Integrated Dial a Ride problem (IDART) with a more data-driven approach and applying process simulation techniques. In the simulation, probability distributions extracted from k-means and facility location will be used as an input. That simulation sample produced from those distributions and at each step solves a new instance of Pickup and Delivery Travelling Salesman Problem (PDTSP). Extended details about PDTSP and PDVRP are also beyond the scope of this deliverable and will be presented in the following documents.

Across those three steps, some parameters can be controlled and allow planners to give to the DRT system the operational characteristics desired. Step 2, for example, allows to control the number of stations, and affect the proximity of passengers to the line. A higher number of stations leads to a lower distance between stops, more passengers satisfied along with higher travel times and increased operational costs. Step 3 allows controlling the vehicle capacity and the frequency. The main purpose of the evaluation step (Step 3) is to create the appropriate metrics that could help planners better understand the dynamics and trade-offs of the system.

A key role here plays the knowledge of origin and destination points. That information can be provided to the method from various data sources like:

- Floating Car Data generated from cars or taxis and have a shape of typical GPS data.
- Mobile network data that are classified as car trip.
- Bluetooth data.
- Survey Data.
- Mobility/General Purpose app that generates mobility data.

The popular K-Means clustering algorithm can be performed using those origin points. This method has some extends that can lead to interesting results. K-Means create approximately 2-D Gaussian distributions (for spatial data) across cluster centres. Those centres can be considered as the nominate DRT stations which line should pick up passengers. Iterating over various cluster centres (that's a parameter for the algorithm), the maximum distance of a possible passenger from the nearest stop can be controlled. Those possible stops can be used as an input to any Facility Location Mixed Integer program that will choose the optimal number and location of DRT stops.

The following model is a typical formulation of Facility Location problem (DS Chen, 2010) with variables:

Minimise $\sum_i \sum_j x_{ij} c_{ij} + \sum_j f_j y_j$

Subject to:

$$\sum_i x_{ij} = d_i \quad \forall \text{ j: 1, …, n} \qquad (1)$$
$$\sum_i x_{ij} \leq \text{ M } y_j \quad \forall \text{ j: 1, …, n} \qquad (2)$$
$$x_{ij} \geq 0 \quad \forall\, i\text{, j: 1, …, n} \qquad (3)$$

where

y_j Takes value 1 if the bus stop will activate with fixed cost f_j and 0 else.
x_{ij} : If i pick up station opened and go to j with$cost\ c_{ij}$.
d_i is the estimated demand for stop i.
M can be replaced with bus capacity

Fixed costs include operational costs, maintenance costs, and possible earnings. More details are beyond the scope of this deliverable and will be provided in WP4 deliverables. After that step and with the use of methods of step 1 it can be extracted the demand distribution of each stop. Constraint (1) ensures that the demand for each stop will be satisfied. Constraint (2) ensures that the capacity of the vehicle cannot be exceeded.

Given the stops and possible destination points of the DRT line, the third step stands as an evaluation module. In fact, the solution of Pickup and Delivery TSP/VRP across many simulation steps used to approximate the distribution of distance and time each cycle route needs. Each instance of the simulation samples from the demand distribution of each stop and use PDTSP to extract the solution. In that way, the output of each iteration contains the total distance travelled, total busses needed, and the total number of station route served. After a few iterations that step can reveal some good approximations of the distribution of those outputs.

2.2.5. Route analysis for ridesharing

Ridesharing refers to the act of sharing a ride with other passengers, preferably with the ones going to the same direction and at the same time window of the day as the driver. The concept is very widely spread worldwide, and the goal is to try to fulfil the capacity of commuting cars, reducing congestion and the environmental footprint (fuel and carbon output, land use, etc.), in addition to reduce the monetary cost for travellers.

In order to analyse the suitability of ride-sharing services in an area it is crucial to know the mobility patterns of the citizens in the area. Transport planners use OD matrices which can represent mobility patterns to an acceptable level, but they have some limitation when we search for decision tools on ride-sharing system planning. It is more accurate to state that along with OD matrices that will reveal patterns between days or maybe hours scale, there is a need for an extra tool that allows visibility and knowledge extraction of trajectories. More specifically, the methods developed, aim to make a descriptive embedding of a route in vectors or some similarity KPIs that could lead to a more meaningful representation via clustering. The final goal of that method is to deploy trip trajectory and extract the most frequent routes along with interesting insights about them.

The trip trajectory is defined as the sequence of points that create a path, including the origin and destination point. The current methodology searches for a definition of similarity between two routes in terms of trajectory. Unsupervised learning techniques are used in this case study to estimate these similarities. The first step is to estimate the similarity of each pair of routes by means of origin and destination. The second step of the study is to identify similar points in the two routes. The matching algorithm presented below searches for similar points used in each pair of routes. More precisely, the algorithm picks one out of two routes and starts iterating over that path. At each iteration, it checks if the current point also belongs to the trajectory of the second route. In case it finds a common node (point in path) add the value of one in similarity measure between those two routes. To ensure that those two routes travel in the same direction, after each positive check (same node in route 1 and route 2), the second route gets sliced from that point and on. That reduction also improves the computational cost as previous points cannot match as route 1 moves towards the destination point.

In Figure 3, a typical high correlated pair of trips presented. Both those trips can easily be served by the same ridesharing line.

Algorithm 1: Matching Similarities.

```
Input:
Route 1 = list
Route 2 = list
Count = 0
For i in Route 1
   If Route1[i] in Route 2
      Current_index = i
       Route 2 = Route2[Current_index:]
      Count += 1
   End if
End for
Return Count.
```

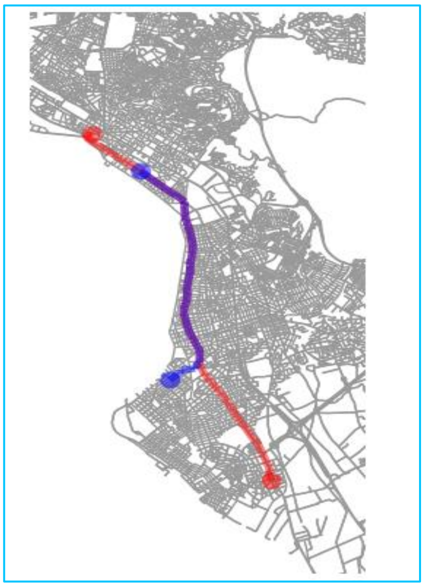


Figure 3. Sample of two Routes with high similarity

In order to turn this information into a feature, it needs to construct the similarity matrix between routes. The similarity matrix is an NxN symmetric matrix that contains the matching nodes across Route i and Route j. A large number of common nodes indicates a higher similarity among routes. Each observation (route) contains a similarity vector that shows how similar is that route compared to every other available routes. Next to that 1xN vector, 4 more features are included containing information for latitude and longitude of origin and destination of the trip. All those values are normalized between 0-1 to avoid bias. Additionally, the similarity vector multiplied by a scalar because it would dominate the cluster similarity measures due to the higher number of features it contains (as the number of routes) compared to the origin and destination features.

X_i = [OriginLat, OriginLon, DestinationLat, DestinationLon] (1x4) + $D_{i.}$(1xN)

Therefore, each trip is treated as an observation with N + 4 features where N is the number of trips/observations.

2.2.5.1. Clustering methods

Various methods that allow to automatically choose the number of clusters are used.

2.2.5.1.1. Automatic clustering

The methodology followed for constructing time depended O-D clusters, is based on creating new zones with different size for each timeslot as is proposed by other works (WERABHAT MUNGTHANYA1, 2019). The steps of the implementation of this procedure are described below:

1. Insert dataset with routes.
2. Split data in time zones.
3. Performs x-means clustering to choose optimal number of clusters.
4. Perform k-means algorithm and extract clusters.
5. Visualization.

2.2.5.1.2. BIC Criterion and X-Means

The BIC criterion (Bayesian Information Criterion) is one of the fundamental methodologies to define the clusters needed. The core idea of the BIC criterion is to measure the volume of uncertainty. As the number of clusters increases the uncertainty (or the information) of the whole system decreases. In a clustering methodology, the basic idea is to create a reasonable number of clusters in order to make a better understanding of the given data. Hence, another parameter that penalizes the expansion of the number of clusters used is defined. In the Bayesian criterion, the penalty factor used increases once the number of clusters increases.

2.2.5.1.3. K-Means

The clustering algorithm that X-means use in order to create clusters is K-Means. It is probably the most famous among clustering algorithms. The core idea is to assign each point in the closest topological cluster. Steps of the process of the implementation of the algorithm are:

1. Choose the number of clusters
2. Initialize centres of clusters randomly in the space of data.
3. Assign each point to the closest centre based on distance measure (Euclidean, Manhattan etc.)
4. For every cluster re-compute the weighted centre.
5. Assign the points on the new centres based on a distance measure.
6. Repeat steps 4-5 until the change of cluster centres seems insignificant.

2.2.5.1.4. Tools and libraries

The tool used in this study is Python in order to analyse received data and produce the results. The libraries are shown on Table 2. The table below, includes the methods used in this study.

Library	Methods	Usage
Pyclustering	kmeans_plusplus_initializer() xmeans()	Computes X-Means clustering. Extract the optimal number of clusters.
Sklearn	KMeans()	Peform K-means clustering.
Pandas	Pandas methods and structures.	Basic library to handle data and manipulate them.
shapely	MultiPoint(),PolygonPatch()	Methods that help to create polygons and find points in each polygon.
Plotly	graph_objs()	Make interactive visualizations.
datetime	Time()	Module to handle time filtering process.
osmnx	graph_from_place(),graph_to_gdfs(), get_nearest_node(),shortest_path()	For extract OSM road network, fit points on network, find paths.

Table 2. Main libraries and modules

2.3. OD matrix comparison and similarity measures

2.3.1. Motivation and objectives

The comparison of Origin-Destination (OD) matrices has important applications such as (K. Behara, Bhaskar, & Chung, 2018): the identification of differences between travel patterns that allow understanding the temporal or seasonal variability of mobility, as well as the influence that special events or incidents can have on it; the estimation of standard or typical OD matrices that allow characterising recurrent travel patterns by clustering similar matrices; or the benchmarking of OD matrix estimation methods, among others.

The classic methods for OD matrix comparison (K. N. S. Behara, Bhaskar, & Chung, 2020a; Djukic, Hoogendoorn, & Van Lint, 2013) have been based on cell to cell comparisons by means of measures such as Root Mean Square Error, Mean Absolute Error Ratio, Entropy, etc. However, these measures simply express the deviations in terms of demand between the compared matrices, and do not take into account the correlations that normally occur between OD pairs and that define what is called the structure of the OD matrix.

In order to address this problem, so-called structural similarity measures have appeared in recent years (K. N. S. Behara et al., 2020a). Unlike the previous ones, these measures implicitly or explicitly define a structure for the OD matrix and then exploit this structure in the comparison. Some examples of these measures are the Mean Structural Similarity Index and its variants, the Wasserstein distance or the Levenshtein distance.

Although the measures of structural similarity proposed so far have proved to be better than the classic measures, the latter present the following problems:

- They are sensitive to certain parameters that do not have a clearly defined value and on certain occasions can be network-specific
- The computation cost of the measure is high, especially for high-dimensional matrices.

With these ideas in mind, and inspired by the definition of structure proposed in (K. N. S. Behara et al., 2020a), in MOMENTUM we have developed a new structural similarity measure for OD matrices that aims to tackle the previous problems by defining a structure for matrices whose comparison can be done in a computationally efficient way, that is, parameter-free and no network specific, and that allows taking advantage of the power of the new High-Performance Computing (HPC) paradigms based on computational graphs (Abadi et al., 2016).

2.3.2. Related work

The previous studies on the comparison measures of OD matrices can be classified into two large categories: similarity measures and structural similarity measures. As for similarity measures, these are usually based on mathematical formulations that calculate the deviation or difference between individual OD pairs. Some examples of proposed measures are the Root Mean Square Error (Ashok & Ben-Akiva, 2002), Mean Absolute Error Ratio (Kim, Kim, & Rilett, 2005), Mean Absolute Percent Error (Cools, Moons, & Wets, 2010), R2 (Tavassoli, Alsger, Hickman, & Mesbah, 2016) or the Entropy Measure (Ros-Roca, Montero, Schneck, & Barceló, 2018) to name but a few. However, as mentioned above, these measures do not take into account the underlying structure of the OD matrices which is due to the correlation between different groups of OD pairs by factors such as generation and attraction of trips, geographical distance or accessibility, among others.

In order to improve the performance of the methods mentioned above, different measures have been proposed that attempt, through different methodologies, to define a structure that allows for a more appropriate comparison between OD matrices. The paper in (Djukic et al., 2013) was the first to propose a similarity measure to address this problem, using the Mean Structural SIMilarity Index, a measure used in computer vision to measure structural degradation between two images. This measure is based on the computation of a measure called Structural Similarity Index (SSIM) between two OD matrices using a sliding window of a pre-established size. The problems presented by this measure are the following: 1) the measure is sensitive to parameters such as the window size and other so-called stability constants, whose variation can give rise to different measurements and their setting can be network-specific; 2) the measure is also sensitive to the order of the rows and columns of the OD matrix, and furthermore, this order cannot guarantee geometric adjacency.

In order to address the problems mentioned for the MSSIM measure, different variants have been proposed, among which the Geographical Statistical Similarity Index (K. N. S. Behara, Bhaskar, & Chung, 2020b), 4D-MSSIM (Van Vuren, T., & Day-Pollard, 2015) and SpSSIM (Jin, Nara, Yang, & Tsou, 2020) stand out. In the first one, the geographical limits of higher-level areas are used to define the size of the sliding window. The second is based on the spatial proximity of the OD pairs (measured by the Euclidean distance between the centroids). And the third one uses a weighted matrix to define spatial adjacency. The first two variants avoid the problem of selecting the window size, but not the last one. Furthermore, they still present the problem of the setting of stability constants, and in addition, the process of identifying the OD pairs that are geographically correlated is network specific.

Another structural similarity measure that we can find in the literature was proposed in (Ruiz de Villa, Casas, & Breen, 2014), and it is based on Wasserstein's distance. When comparing two OD matrices, this measure can be understood as the minimum travel time required to allocate the trips between the OD pairs in matrix A using a trip allocation that is compatible with the structure of matrix B. Although this measure allows capturing the differences in the distribution of trips between two matrices, its main problem lies in its high computational cost, since it is based on the resolution of an optimization problem whose dimensionality grows according to a four-degree polynomial order.

Finally, we have the measure recently presented in work by (K. N. S. Behara et al., 2020a). This paper is the first to explicitly conceptualise different elements that would define the structure of an OD matrix prior to attempt a similarity assessment. Concretely, the authors define the structure by means of two concepts: 1) the skeleton or structure of the OD matrix, which consists of the preference of the destinations for each origin (the order of the destinations according to the number of trips); and 2) the mass, which corresponds to the OD flows ordered according to the structure defined by the skeleton matrix. Following this definition of structure, the authors establish the structural similarity between two matrices by means of what they call normalized Levenshtein distance for OD matrices (NLOD) which, as its name indicates, it is based on the well-known Levenshtein distance. Although this measure solves several of the problems mentioned above, such as sensitivity to parameters (it is parameter-free), it has two main shortcomings. Firstly, its computational cost, since it presents an $O(n^3)$ complexity, which can make it slow for large OD matrix sizes. And secondly, the definition of structure they propose is sensitive to noise or small variations. Being more specific, in the case that for a certain origin there are destinations with a similar number of trips, small variations in their values can cause large changes in the order of the destinations and, in turn, in the results of the measurement.

2.3.3. Description of the new Structural Similarity Measure for OD matrices

The main motivations behind the conceptual design of the structural similarity measure for OD matrix comparison designed in MOMENTUM have been the following:

- **The design of the measure should be robust to noise** so that variations due to small fluctuations in mobility or small errors in estimating OD flows do not lead to very different values of similarity.
- **The design of the measure should be parameter-free and network agnostic** to improve its robustness and facilitate its applicability for the comparison of OD matrices in different locations and for different transport modes and data sources. By network-agnostic we refer to the fact that neither the measurement nor its application depends on parameters or other factors whose adjustment varies according to the specific characteristics of the transmission network for which the OD matrices have been defined. This facilitates their applicability, as indicated above.
- **The definition of a concept of OD-matrix structure whose comparison is computationally efficient**, so that a large-scale comparison of OD-matrices can be addressed. In this way, it would be possible to compare a larger number of OD matrices. The increasing availability of fine-grained and longitudinal data sources for describing travel demand suggests that this ability can help transport modellers. It enables the comparison of more disaggregated OD matrices (e.g., at hourly or time-slot level), more efficient estimation of representative OD matrices and the application of Machine Learning techniques to predict or estimate OD matrices since many of these techniques are usually based on the iterative application of a measure of error, distance or similarity.
- **This design should be easily deployable in new HPC paradigms based on computational graphs** such as TensorFlow or PyTorch. These paradigms allow very efficient operations to be carried out with multi-dimensional arrays, called tensors. And what is more important, they allow taking advantage in a very simple and transparent way of the high computing performance of current hardware devices such as GPUs, as well as great facilitation for the parallelization of calculations.

Having said this, the general features of the design of the structural similarity measure are described below:

- **The structure of an OD matrix is defined by the preference/arrangement of both destinations for each origin, and origins for each destination**. Unlike the approach proposed in (K. N. S. Behara et al., 2020a), our concept of structure is not based only on the order of the destinations for each origin (or origins for each destination), with the aim of making it less sensitive to noise. Furthermore, it is not based on a pre-processing of the original OD Matrix, which leads to a more efficient concept of OD Matrix structure.
- **The design of the structural similarity measure has no parameters in its definition to facilitate its use and applicability**. The measurement internally differentiates between elements similar to what Behara

et al. call structure and mass. Although a weighting could be given to each of these elements, we have chosen to define a mechanism for its self-regulation. This makes the measure, in principle, network agnostic. However, although the first results point in this direction, to validate this fact, as we will comment later on, further experiments will be carried out in the next stages of the project.

- **The algorithmic complexity of the calculation of the structural similarity between two OD matrices is O(n^2)**
- **The definition of the measurement has been made based on operations with multi-dimensional arrays or tensors** and has been implemented using TensorFlow 2.0. This has allowed performing a comparison between two OD matrices of size 208x208 in approximately 0.12 seconds, using Google Colab and a virtual machine with a Tesla V100-SXM2-16GB GPU (compute capability 7).

2.4. Shared mobility user diaries

2.4.1. Motivation and objectives

The service deployed by each shared mobility operator has its own particularities and conditions, that can range from the vehicle provision model (dock-based or free-floating), the service area or geofence, the fare schemes or even the vehicle provided (car sharing, bike sharing, etc.). Indeed, data collection procedures depend on the goals and interests of each provider, which determines the type of data collected at the operation phase. In sum, the process of data collection and storage is unique and independent for every single provider, which results in large differences observable in the service data obtained by each operator. The recent implementation of certain standards for sharing mobility data from these new services (Zach, 2019), such as the Mobility Data Specification (MDS), may contribute to reducing these differences. However, these standards are not yet widely implemented.

This current lack of standardization complicates any kind of efficient analysis since routines and procedures designed for the data of a service provider will surely not be valid for the data of any other. Apart from promoting the adoption of standards such as MDS, transport practitioners require standards to process the already available heterogeneous data from different services.

In this light, a generic approach to loading, homogenizing and modelling data from shared mobility operators has been explored. The approach is based on the creation of an entity diary as the standard format. In shared mobility systems, the entities of interest can be either users or vehicles. An entity diary contains all the trips performed by an entity during a fixed period of time, typically a day, along with other relevant details of the entity during the period.

2.4.2. Diary structure

2.4.2.1. User diaries

A user diary aggregates all the trips performed by a shared mobility user for a given period of time in JSON-like format. The diary contains at least the following fields:

- Date
- User identifier
- Trips information, containing all the trips the indicated user performed during the stated date.

In addition, user characteristics, like age or gender, or any other time period specific characteristics, like special events or weather conditions during the period of the diary can be included at the topmost level of the diary of each user. The diary register contains a list of trip entries, which are JSON-encoded and necessarily contain for each trip the initial and ending times and position coordinates recorded for the trip as well as any other optional parameters that can be included in the service, such as battery levels, station-based positions or specific billing details.

This user diary definition simplifies the compulsory fields to the minimum needed parameters to define each trip, as well as the collection of trips per user and day. Furthermore, the use of optional parameters under a common definition scheme enables more complex analyses based on user diaries that consider optional fields for those services that provide them.

The support for optional parameters is the main reason for the use of a JSON-like data format since it deals better with non-required parameters than classical formats like tabular data. The JSON format is also very flexible and accepts more complex data structures, such as lists or other JSON objects that are useful to store the details of each trip.

2.4.2.2. Vehicle diaries

Vehicle diaries are an entity-based alternative to user diaries. Conceptually they are the same as user diaries but changing the trip aggregation criteria: instead of aggregating all the trips performed by a shared mobility user, vehicle diaries aggregate all the trips performed in the same shared vehicle within the service. For the project, user diaries have been used, but vehicle diaries could be derived similarly.

2.4.3. General user diary generator implementation

User diaries are stored by day, so for each day in the period of study, a file is stored in the storage file system. Each file contains a line for the JSON-encoded user diary of every user that performs at least a trip in the service during the date of the diary.

In the scope of the MOMENTUM project, a general object-oriented framework for shared mobility user diary generation has been implemented in Python programming language. The framework defines two basic data structures to represent a trip in shared mobility services and an entity diary structure. Each of these structures implemented as classes, provide methods to transform them into a JSON-like data structure to be stored in a file.

Additionally, each of the structures provides an interface for data loading to be implemented differently for the data sources of each shared mobility service. Actually, data loading into these structures is the only routine to be implemented for each new data source, leaving the process of diary generation to methods implemented by the generic data structures.

2.4.4. Generation of OD matrices from user diaries

Once the basic format of user diaries is available, OD matrices are the most relevant product to be obtained from them. Basically, OD matrices can be obtained from user diaries by assigning trip starting times to their initial hour and the initial and final coordinate positions to origin and destination zones according to a study zoning scheme that shall be provided by the user.

2.5. Shared mobility user profiling

2.5.1. Motivation and objectives

The recent emergence of shared mobility services implies that there is still limited information about the profile of their users and the services adoption patterns. Among the most common sources for analysing these aspects, the following can be highlighted:

- **Registration data from service users**. Many services require their users to provide certain data about them, such as home location or age. The fields registered and the level of detail usually depends on the characteristics of each service. For instance, car-sharing and moto sharing systems require the validation of the user's driver license, which includes information about the age of the user. All this information may not be available with the desired level of detail for the analysis due to privacy constraints, and it is also usually the case that the information is not complete for all users. In any case, some relevant attributes (e.g. income, car ownership, …) are hardly ever reported in these registration data.

- **Surveys handed to users**. Operators usually collect data from surveys, taking advantage of the mobile app that is needed for using the service. The data collected from these surveys suffer from the limitations of this methodology: limited sample size, response biases, etc.
- **Household surveys**. They provide an alternative to these service surveys and enable a comparison of the user profiles vs non-user profiles. The most recent household surveys already include the new modes, as can be seen in Section 3 for some of the cities hosting MOMENTUM case studies. However, household surveys are very resource consuming and therefore are not frequently updated. In addition, the fact that shared mobility services still account for a limited proportion of the overall mobility implies that the available sample of shared mobility users in these surveys is small.

The aforementioned sources try to retrieve the characteristics of each individual by asking them to provide information. An alternative approach based on data fusion can be of interest. Certain features of the users can be very related to the home location, as major sociodemographic studies are typically available on a per-home basis, generally at the census tract, district or municipality level. Hence, if the home location is known, the limitations entailed by registration data and surveys can be partially solved. However, as explained above, it is usually the case that the home location information is not available for all users or it is available at a spatial level (e.g. zip code) that does not match the resolution of the sociodemographic data that is required to attribute the target features.

As a consequence, MOMENTUM has explored the development of methods that (1) enable home location inference from the operation data that is passively collected by service providers to complete the information provided by user registration; and (2) attribute an income level to the shared mobility users according to their home location.

2.5.2. User home inference

The availability of operation data, recording all trips made by each user, is an opportunity to complement the home location data collected in the user registration process. This can be helpful in two ways: (1) disaggregate the available home location data that can be shared to a level that allows statistical matching with sociodemographic data, (2) identify the home location of those users with no information in the registration data. Mobility data obtained from shared services could be useful to derive the most probable home location of a user, given their trips in shared mobility modes. The present section reports the different methodologies developed to infer home location from mobility data from different shared mobility services. Two cases can be differentiated: (1) cross-sectional home inference methods, (2) longitudinal home inference methods. The methods are applied to the operation data from shared mobility services available for the project. Their applicability can be validated against the information available for all or part of the service users via their registration data.

2.5.2.1. Cross-sectional home inference methods

Cross-sectional home inference methods are the ones applicable to those operation datasets that do not employ the same user identifier through all the periods to be analysed. Some shared mobility providers re-anonymise their datasets with a given periodicity (e.g. each day), which makes not possible to characterise longitudinal travel patterns at individual level. This section describes those home inference methods that have been developed to be used over a single user diary for an entire day with no recurrent information through time.

2.5.2.1.1. Home inference based on morning hour range

This method is based on the travel behaviour in the first hours of the day. The underlying hypothesis is that the origin of the trip made during certain hours in the morning is likely to be close to the user home location for a significant proportion of users.

Hence, this algorithm selects as home location the origin location of the first trip of a user that takes place between 6AM and 10AM. The algorithm follows the next sequence of rules:

- Taking as input any user diary. For each user:
- Look for the first trip in the range **6AM-10 AM**:
  - Set that location as candidate home location.
  - If no trip is available during the hour range, the home location is set to “Empty”.

2.5.2.1.2. Home inference based on roundtrips

This method is based on the observed roundtrips for each user. The underlying hypothesis is that those locations where the user base its roundtrips may correspond to the home location. In this context, a roundtrip is a pair of trips that go from a point A to a point B and afterwards comes back from B to A. The home location would be then point A where the trip tuple starts and comes back to. The flow of events for this home inference method is the following:

- Taking as input any user diary. For each day:
  - If user has performed more than one trip:
    - Retrieve all possible roundtrips in the diary.
    - The home location is set to the **most frequent roundtrip location** for that day.
  - If the user has a single trip record, the morning hour range method is used instead:
    - Take the first trip in **the morning range 6AM -10 AM** as home location.

2.5.2.2. Longitudinal home inference methods

The main limitation of the home inference methods defined above is that they are constrained to the information observed in a single day, which neglects the longitudinal information provided by regular user trips and changes according to the day of the week. If longitudinal data is available, most frequent locations should be analysed, since it is possible than one of the most frequent locations of each user is their home location.

However, there are several ways to assess the location appearance frequency. For that purpose, the following list provides a set of methodologies that targets recurrent locations of the user:

1. **Trip location recurrence**: the most frequent location in each user’s trips will be selected as their home location.
2. **Origin trip location recurrence**: assuming most trips could be originated from home, the most frequent origin location in each user's trips will be selected as their home location.
3. **Destination trip location recurrence**: opposite to the previous method, this one assumes that most trips end up destined to home and therefore, the most frequent destination location in user’s trips will be selected as their home location. This tries to take into account the existence of daily triangular travel chains where one of the trips may be based on work location rather than on home location.
4. **Trip location recurrence in morning range (6AM-10AM)**: following the same hypothesis than in the cross-sectional case, the most frequent origin zipcode in the morning range between 6 AM and 10 AM is assigned as home location.

5. **Origin and destination most connected trip locations**: home location may be characterised by being the location from/to a larger range of different origins/destinations are observed. Thus, the most connected locations, those being the origin or destination from/to most different locations are set as home location candidates.

2.5.2.3. Machine learning powered home inference

So far, the methodologies proposed are based on heuristic rules regarding well-known behaviours of travellers, such as trips originated on certain hours from home or places where users end up coming back to. Despite, aiming at a broader inspection at home location inference from shared mobility data, a machine learning classification formulation has been proposed for home inference. This also enables a detailed assessment of the potential of the heuristic methods, by looking at the accuracy of the classification with regard to the declared home location information in the registration data.

The classification problem consists of a binary classifier that determines whether a location is the home location of a given user taking as input the collection of ranking variables obtained following the previously proposed methodologies. The list of features is the following:

- Position in the ordered list of locations that a user travels to or from.
- Position in the ordered list of locations that a user travels from (origins).
- Position in the ordered list of locations that a user travels to (destinations).
- Position in the ordered list of origin locations that connect with more locations.
- Position in the ordered list of destination locations that connect with more locations.
- Position in the ordered list of locations that most frequently appear in roundtrip.
- Binary value indicating whether the zipcode appears in the morning range.
- Total number of zip codes for the user.
- The same positions in the rankings (zipcode, origin, destination, connections and roundtrips) for the best scoring neighbour zip code each time (the best position in the zipcode ranking among neighbours, etc.)

The model selected for this task is a decision tree model, mainly due to its design and the possibility of printing the tree and the decisions it takes to reach the classification goal. This way it is possible to recover a set of data-driven rules that consider different dimensions to determine home location of a given user.

2.5.3. User income assignment

One of the aspects that cities highlight as relevant when assessing the impact of shared mobility services is the affordability of these new transport modes. This is one of the dimensions for the inclusiveness assessment of emerging mobility solutions, together with others such as age bias. Hence, a method has been developed to attribute income levels to shared mobility users based on the home location information and the income spatial distribution data. This method can be applied either to the declared home location or to the inferred home location following the algorithms explained in the previous section and has to be adapted to the information available for each service.

In this case, two different data availability cases were identified: (1) a customer database available with the home location of each registered user, or (2) home location of the user performing each trip included in the trip data. The latter is related to those services that do not provide longitudinal information since they re-anonymise the user identifiers every day. Hence, the data provide the home location information for a 'user-day' rather than for a 'user'. Nevertheless, the total number of users is usually known as it is less sensitive in terms of privacy.

2.5.3.1. Customer database with home location information available

For each user appearing in the customer database, the income level of the administrative unit declared as its home location is assigned to the user.

2.5.3.2. Trip data with home location information available

When the home location information is available for 'user-days', extracting the entire user distribution within the system is a non-trivial task. Since the user IDs in the trips made by a single person are not consistent throughout the analysed periods, it is not possible to simply recollect every new user ID observed. In this light, it is required to develop a methodology to derive a large enough sample of users that live in each administrative unit from daily trip registers observed during all the time of study. The workflow is the following:

- Obtain the total distributions of user home locations by counting the number of times each administrative unit appears each day as a user home location.
- Compute the average number of users each administrative unit is reported as home location for every day in the period.
- Obtain an expansion factor by dividing the total number of users by the total number of observed users along the analysed periods (sum of the total count of administrative units) and multiply each unit count by them.
- Recreate the income distribution by appending to the distribution the income at each administrative unit as many times as users have been provided.

Following this methodology, it is possible to obtain a representative distribution of the user location (and thus, income) distribution. This can be validated by replicating the process in those services where customer database information is provided.

3. Data analyses and validation of new methodologies

3.1. Madrid Case Study

3.1.1. Data sources

The main data sources used in Madrid case study are shown in Table 3. A more detailed description of the data sources can be found in deliverable D3.1.

Data source	Data source	Data source
Income Statistics	SD_002	Average income figures for the population living in Spain at the census tract level.
Mobility household survey	TD_006	Regional mobility household survey conducted in 2018.
Moto sharing OD matrices from operation data	TD_009	Hourly origin-destination trip matrices obtained from moto sharing operating data, quantifying service demand. The data used for the analysis covers October 2018, December 2018, February 2019, April 2019, June 2019, July 2019 and October 2019, December 2019 and February 2020.
Overall mobility OD matrices from mobile phone data	TD_007	Hourly origin-destination trip matrices obtained from mobile phone data, quantifying overall mobility demand (regardless of the mode). The data used for the analysis covers October 2018, February 2019, April 2019, June 2019, July 2019 and October 2019.
Station-based bike sharing OD matrices from operation data	TD_003	Hourly origin-destination trip matrices obtained from bike sharing operating data, quantifying service demand. The data used for the analysis covers October 2018, December 2018, February 2019, April 2019, June 2019, July 2019 and October 2019, December 2019 and February 2020.

Table 3. Main data sources used in Madrid case study

The Madrid region "Encuesta Domiciliaria de Movilidad" (EDM) (TD_006) is a household mobility survey developed by the Madrid regional government during 2018 to learn how does Madrid population travel within the region on a working day. The proposed questionnaire inquired about the number of trips and mobility choices of travellers living within the region in a recent working day. User information included sociodemographic characteristics along with other relevant mobility-related issues (public transport card, available vehicle, etc.). The survey took place from February 2018 to June 2018 and performed a personal interview to a random sample of inhabitants in the region through two channels: presential interviews to all the members of up to 13,009 households (36,653 people) and individual phone interviews to up to 50,412 people. As a result of this random sample strategy, the survey provides for each household, individual and trip an expansion factor to generate an OD matrix estimation based on the correct representation of each individual from the sample into the entire population. The data is shipped under four aggregations: households, people, trips and trip legs. Each aggregation has a different ID that serves as a linker to be grouped with the rest of aggregations. The combination of IDs uniquely identifies the elements in the lowest level of aggregation; for instance, household ID and individual ID together can uniquely identify every traveller in the people view, whereas the traveller ID by itself is repeated through households. More details on the methodology can be found in CRTM (2019).

The methodology used for extracting OD matrices from mobile network data (TD_007) is the one reflected in previous works such as Bassolas et al. (2019). The raw data include (i) all the anonymised call detail records (CDRs) stored by the mobile operator for billing purposes; (ii) the cell tower network map of the operator; and (iii) basic sociodemographic information associated to each anonymous mobile phone. CDRs are produced every time a mobile phone interacts with the network through a voice call, a text message or an Internet data connection. Among other information, each CDR contains an identifier of the user, a timestamp and a location. The temporal granularity of CDRs depends on the mobile phone frequency of use, but it usually provides a register for each user every 15-30 minutes. The geographical granularity of the data depends on the number of cell towers (in metropolitan areas such as Madrid, this typically ranges from 100 to 500 metres). Given the market share of the operator, the sample size of this data source is larger than the 20% of the Spanish population. Each device is conceptualised as an agent constituting the sample. The methodology is based on the generation of activity-travel diaries for each agent with enough registers to characterise their mobility. The longitudinal analysis of the activity-travel diaries leads to the identification of frequent activities, such as home and work location. Home location identification enables sample expansion by factors relating to the sample size available at each administrative unit and the resident population. Some basic sociodemographic information is linked to each device through the contract information (age, gender), so this expansion can be done by population strata. Once the activity-travel diaries of each agent are expanded by the corresponding factor, the OD matrices are trivially generated aggregating the information at the required spatial (study zoning) and temporal (periods) levels. In this case, the study zoning used was the one used by the transport authority CRTM for the aforementioned survey.

The BiciMAD service is a station-based public sharing bicycle system that operates within the city of Madrid. The operation data available makes it possible to extract OD matrices at any required level of aggregation (TD_003). It relies on electric bicycles available through a network of fixed stations distributed throughout the city where they are charged and maintained. The service is operated by EMT Madrid, the city's public transport company which is also in charge of bus services and offers discounts to public transport users that possess the monthly public transport card.

Similarly, Muving is a public motorcycle sharing service based on floating motorcycles within the city of Madrid. The service is run by a private company and is deployed in another five cities in Spain. At the time of writing, Muving service does no longer operate in Madrid, as they sold their business to a competitor. In any case, the service was fully operational in the city during the months of study and OD matrices can be extracted from the data at any required spatial level (TD_009).

User income is a relevant variable for the characterization of users within their socio-economical dimension. The National Institute of Statistics of Spain together with the treasury has recently published the Spanish "Household Income Distribution Atlas", which provides figures of the average income of the population at the census tract level (SD_002). Hence, it is possible to use the home location of shared mobility users to infer their income distribution and compare it between services as well as with the general population distribution.

3.1.2. Comparison of household survey data and mobile network data

3.1.2.1. Motivation

Regional and metropolitan transport authorities regularly undertake wide transport surveys among their inhabitants in order gain insights on the mobility patterns in their areas, which may be used to adapt and prepare the public transport systems to the observed demand and mobility trends. In the region of Madrid, the household mobility survey inquiries about the mobility patterns of a panel sample of the region's inhabitants which can be expanded to the entire region population.

However, it is well known that mobility surveys have certain limitations. They are usually too expensive to conduct them frequently, so the information is usually outdated. The questions asked and how each respondent remembers their mobility for the day in question introduces several biases. With the advent of passively collected big data sources, such as mobile network data, new tools to understand mobility patterns have appeared, providing new possibilities to estimate overall mobility patterns.

This section presents a comparison between the mobility OD matrix extracted from the Madrid's region 2018 mobility survey and the OD matrix extracted from mobile network data to unveil and understand the advantages and limitations of each approach as well as to validate that the estimations from both approaches are similar enough.

3.1.2.2. Methodology

This section develops a comparative validation of the OD matrix estimation approaches based on mobile network data and mobility household survey data. For that purpose, the OD matrix resulting from expanding the 2018 Madrid regional mobility survey (EDM) will be compared to an average working day mobile network data OD matrix computed as an average of the OD matrices from the 18th, 19th and 20th of February 2020. The mobility household survey tries to represent a standard working day in non-summer period, and February is usually employed a standard month for these analyses.

The OD matrix from mobile network data has been generated filtering out those trips that are not captured by the survey, to enable a more accurate comparison. Concisely, the filtering criteria have been to remove the following trips:

- Trips of less than 500 meters distance.
- Professional trips: those performed as a result of working activity, such as delivery or taxi-driver trips.
- Travellers not living in the region of Madrid.

The OD matrix generated from mobile network data is segmented according to four trip purposes: work (W), home (H), other frequent activities (O) and non-frequent activities (NF). This segmentation comes from the individual analysis of the frequent locations for each user part of the mobile network data sample. One of the drawbacks of this data source is that it does not provide as detailed information on trip purpose as surveys do. It has to be noted that it is complex to correspond typical trip purposes reported to surveys to these four categories, given the indetermination between O and NF purposes in relation to many non-work-related trips. However, to evaluate the differences between both categorisations in the datasets, a tentative correspondence has been explored:

1. Home → H
2. Work → W
3. Work errand → NF
4. Study → W
5. Shopping → NF
6. Doctor's appointment → NF
7. Going with someone → NF

8. Leisure → NF
9. Sports/Walking → O
10. Personal errand) → NF
11. Going to another home (second residency) → NF
12. Other → NF

After this pre-processing, survey and mobile network data-based OD matrices are systematically compared in global and relative terms, paying special attention to common matrix segmentations, such as gender, trip purposes or trip distances.

3.1.2.3. Overall trip volumes

The trip aggregation of the EDM matrix accounts for a total of 15.8M trips whereas the mobile network data OD matrix provides an estimation of 16.5 M trips. This difference is coherent with Nommon's experience in the field, where most clients agree surveys tend to underestimate overall trip volumes. This can be due to a systematic elusion of certain non-work-related trips.

Furthermore, we obtain the number of trips per person by dividing overall trip volumes by the population of the region, of approximately 6.5M people. Using this number, the number of trips per person for the EDM is 2.43 compared to the 2.53 trips provided by the mobile network data matrix, which is in line with the underestimation observed in total trips from both approaches.

3.1.2.4. Sociodemographic profiles

3.1.2.4.1. Gender

Both matrices provide segmentation information of the trips, so they can be directly compared. The following barchart in Figure 4 provides the gender share in each of the surveys.

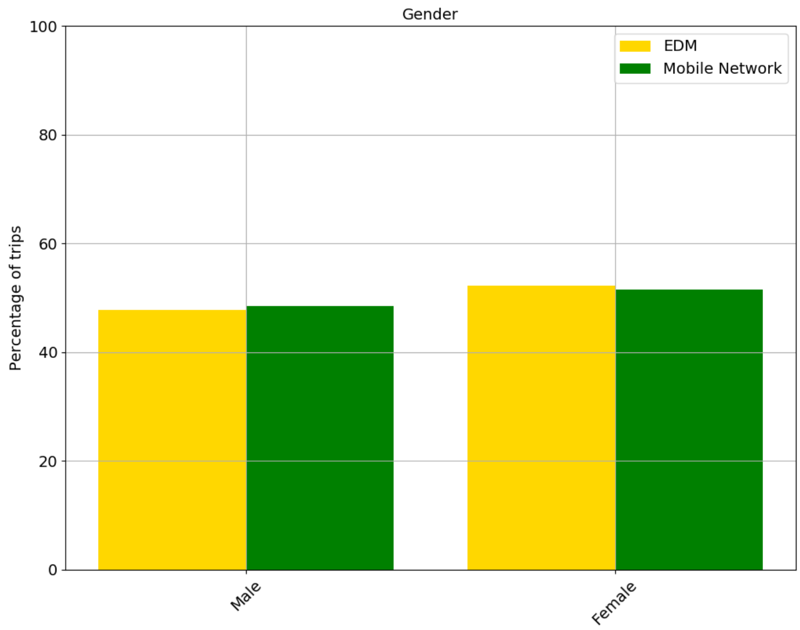


Figure 4. Comparison of OD matrix trip share by gender.

In spite of a very subtle difference in the exact amount, the number of male and female travellers reported in both matrices are consistent and very close.

3.1.2.4.2. Age

Similarly, EDM provides the age of each traveller, while the mobile network data-based OD matrix provides segmentation by age ranges, so it is possible to aggregate travellers in the survey into the mobile network data age ranges. Figure 5 displays the barchart comparison of both approaches by age ranges.

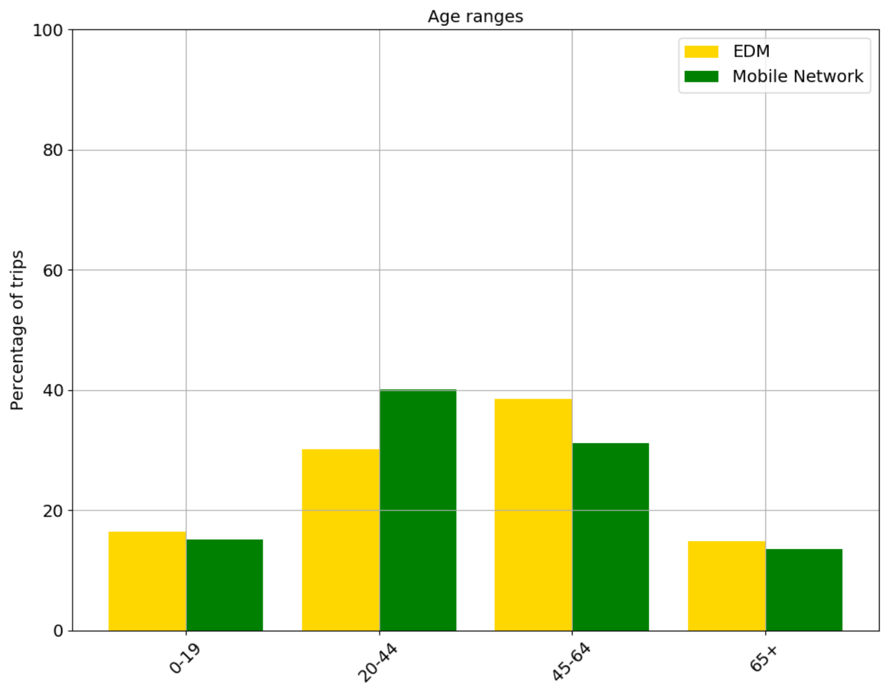


Figure 5. Comparison of trip shares by age group.

In this case, there is a noticeable difference between approaches in the groups 20-44 and 45-64 of almost ten points. The mobile network data matrix seems to be detecting more trips in the age range 20-44 while the EDM has larger shares in the rest of age groups, especially in the 45-64 age group. The most probable cause for this imbalance is due to border effects along with some imprecisions with the ages of the people registered as phone owners.

3.1.2.5. Trip distribution patterns

Trip distribution patterns can be compared by looking at the trip distance distribution (aggregate comparison) and at each cell of the matrix (OD pair comparison).

The survey OD matrix has been obtained following the study zoning scheme of the transport authority of the region (Consorcio Regional de Transportes de Madrid, CRTM) which consists on a large granular study zoning scheme that divides the region into 1,259 zones. Mobile network data-based OD matrices can be computed for any given zoning, and since the CRTM one is publicly available, mobile network data-based OD matrices have been computed with it. Additionally, the study zoning provided by CRTM supports three different aggregations that consistently yield larger zones up a ring-based scheme consistent on four huge zones centred at the city of Madrid.

OD pair values of any two OD matrix estimations for the region should be consistent, following similar trip flow and structure patterns. To verify this, trip volumes are compared at each level of aggregation by means of the Pearson correlation coefficient between both approximations aligned by OD pair.

3.1.2.5.1. Trip distance distribution

The distance for each of the trips in both approaches has been measured. In the case of the mobile network data matrix, the distance is computed as the haversine distance from the centroid of the origin zone to the centroid of the destination zone, whereas the EDM does not report how distances are computed (the information is included in the microdata registers). For comparison purposes, distances have been grouped in ranges and displayed in Figure 6 below.

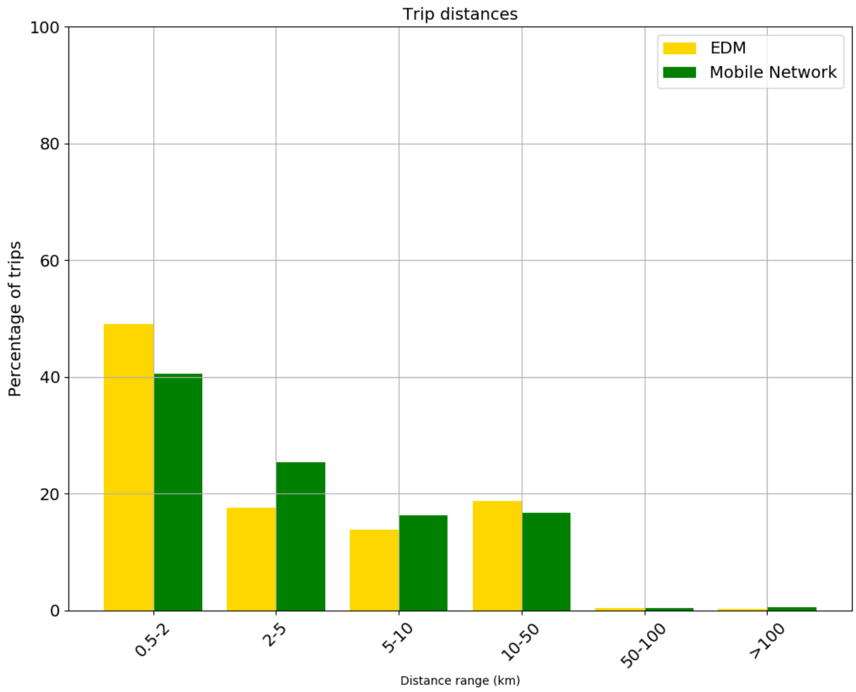


Figure 6. Trip shares by distance range groups.

Figure 6 shows that both approaches seem comparable, having mobile network data matrix slightly more trips with distances between 2 and 50 km and the EDM having more trips reported as less than 2 km. In the rural areas of the region, some of the trips in this range 0.5-2km may be not visible for the mobile network data, as they would happen within a single coverage area.

3.1.2.5.2. OD pair comparison: 1259 zones

Figure 7 shows a visual comparison between both OD matrices at the most disaggregated level. It can be seen how the household survey have a much larger number of 'zero cells', given its smaller sample size. Figure 8 displays the correlation plot of both OD matrices for the most granular study zoning scheme of those provided by the regional transport authority.

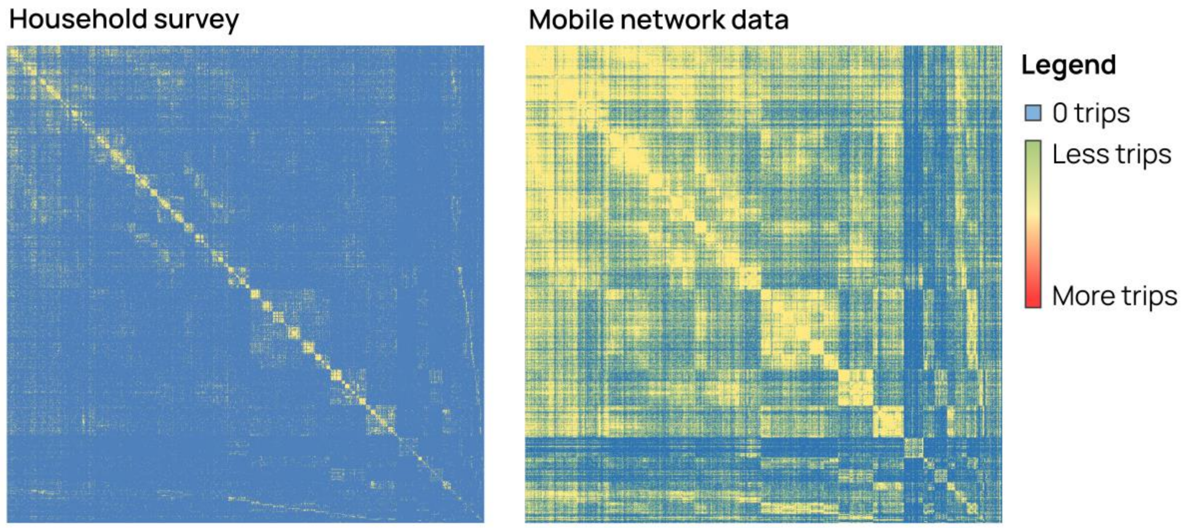


Figure 7 Visual comparison of mobile network data and EDM OD matrices for 1259 transport zones.

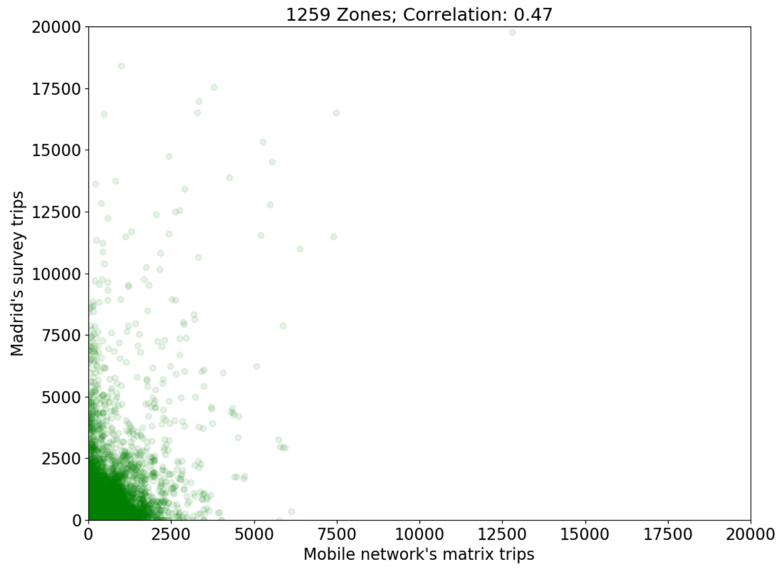


Figure 8. Correlation between mobile network data and EDM OD matrices for 1259 transport zones.

The figure shows that this scheme results in a moderate correlation value of nearly 0.5, even though the plot displays a very noisy observation with no clear linear trend. First, even though mobility patterns are very similar at aggregated level between the periods analysed for both data sources (standard months in 2018 and February 2020), it is likely that differences arise at micro level. Second, at this highly disaggregated level, the sampling errors of both methodologies that are applicable to each OD pair can compensate or amplify each other. Third, the measurement errors of both methodologies can also compensate or amplify each other. In the case of surveys, those OD pairs with a higher proportion of non-work-related trips are likely to be underestimated. In the case of mobile network data, the activity assignment from coverage area to study zoning can be subject to errors at highly disaggregated levels. In general, mobile phone users can be located into the area of coverage of an antenna whenever they are connected to it. Such areas of coverage do not follow standard zoning systems and, therefore, the existence of antenna areas shared by two or more transport zones can be frequent and introduce noise in the areas, especially when zones are comparable in size to the area of coverage of the antennas. This is partially alleviated by using weighted assignment procedures according to the land uses present in the coverage area.

3.1.2.5.3. OD pair comparison: 208 zones

Figure 9 shows a visual comparison between both OD matrices using the study zoning of 208 zones. The structure is fairly similar between both matrices, but the household survey fails to sample trips in the OD pairs with less trips. Figure 10 shows the correspondence between the trips of both OD matrices for the study zoning of 208 zones.

In this case, the expected linear pattern is now observed in the figure, showing how larger zones reduce the distortions introduced by comparable sizes and therefore increases the obtained correlation value up to 0.87 which denotes a strong correlation that confirms the similarity of both approaches. The sampling and measurement errors explained above are likely to be reduced for both data sources at this level.

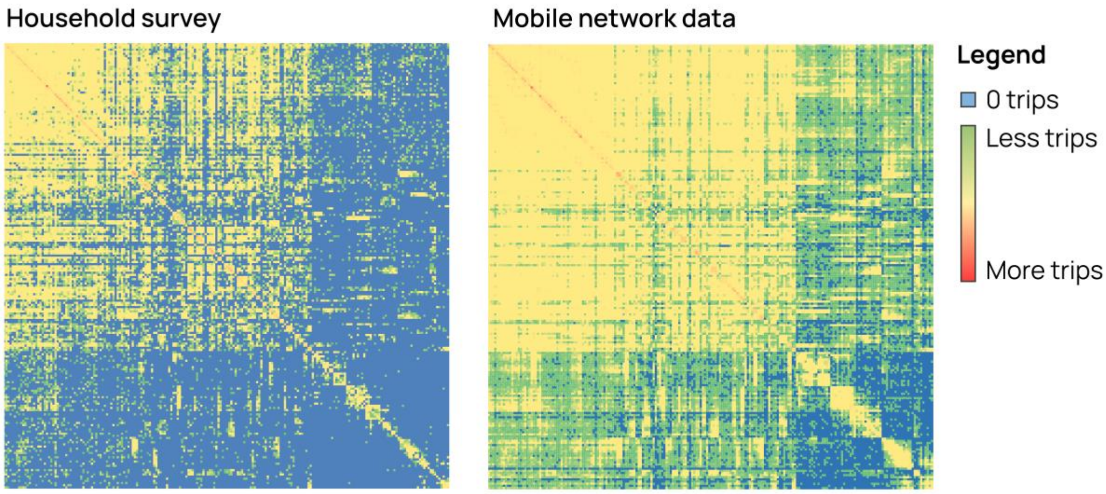


Figure 9. Visual comparison of mobile network data and EDM OD matrices for 208 transport zones.

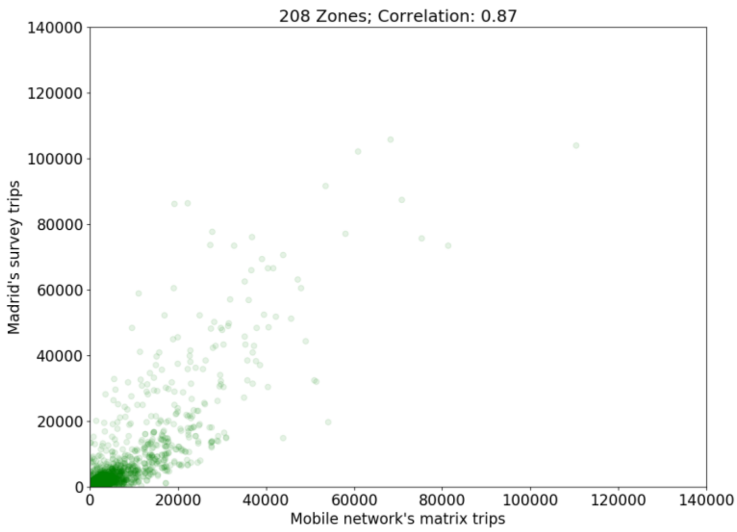


Figure 10. Correlation between mobile network data and EDM OD matrices for 208 transport zones.

3.1.2.5.4. OD pair comparison: 89 zones

Figure 11 presents the correlation plot of both matrices for the aggregated zoning consisting of 86 transport zones aggregated from previous schemes.

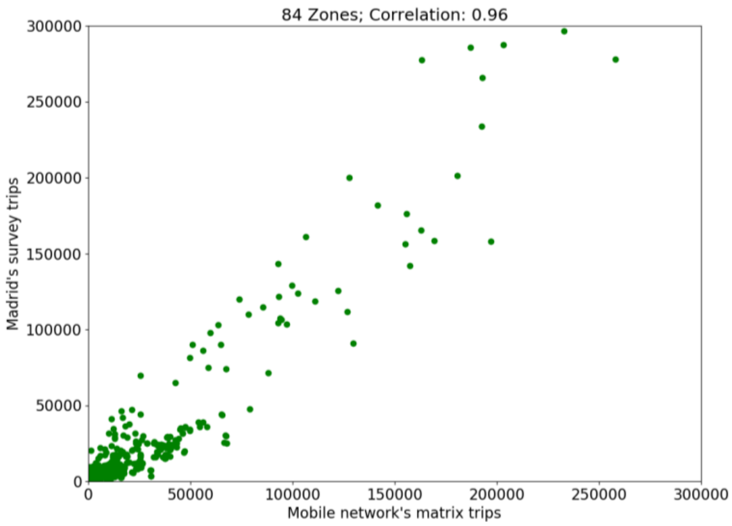


Figure 11. Correlation between mobile network data and EDM OD matrices for 89 transport zones.

In this case, the figure shows how the linear pattern is more noticeable along with a significant increase in correlation value to 0.96, very close to 1, which is a perfect correlation. Apart from the improvement, the trend confirmation further sustains that the trips estimated by both approaches are consistent.

3.1.2.5.5. OD pair comparison: 4 transport rings

Finally, Figure 12 displays the correlation for the four broader rings defined. The rings have been defined by CRTM and are centred in the city of Madrid, being the first ring (Almendra) the circle covering Madrid city centre, the second ring (Periferia) the suburbs of the city, the third ring (Metropolitana) the metropolitan area and the last ring (Regional) all the area in the region outside of the city and its metropolitan area.

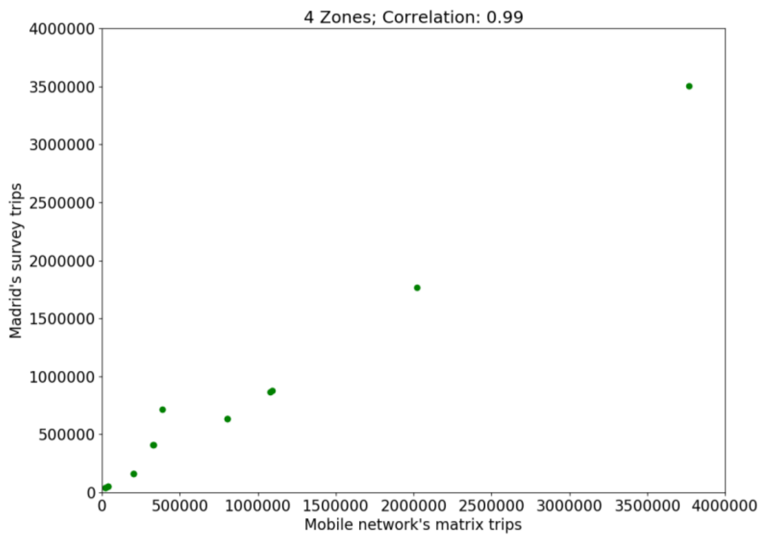


Figure 12. Correlation between mobile network data and EDM OD matrices for 4 transport zones (rings).

In this case, the points in the graph draw an almost perfect straight line, which denotes the practically perfect correlation value observed (0.99). As a result, it can be concluded that both matrices are very similar in terms of flows, especially those flows between large and medium-sized transport zones.

For further details, Table 4. Relation between the shares of mobile network data and EDM OD matrices for each OD pair of the four transport rings study zoning. reports the share of trips that move between transport rings for the survey and Nommon matrices including both inbound and outbound mobility flows. For instance, the Almendra-Periferia case includes all trips from “Almendra” to “Periferia” and all trips from “Periferia” to “Almendra”:

Ring combination	Nommon share (%)	EDM share (%)
Almendra-Almendra	12.34	11.25
Almendra-Periferia	13.41	11.12
Almendra-Metropolitana	4.04	5.21
Almendra-Regional	0.27	0.52
Periferia-Periferia	22.95	22.31
Periferia-Metropolitana	10	8.07
Periferia-Regional	0.47	0.65
Metropolitana-Metropolitana	31.66	34.24
Metropolitana-Regional	2.5	2.08
Regional-Regional	2.35	4.54

Table 4. Relation between the shares of mobile network data and EDM OD matrices for each OD pair of the four transport rings study zoning.

The table displays a very high correspondence level in amounts, especially in those combinations including larger amounts of the regional total. The correlation between the shares of both OD matrices is 0.988, which is very high and in line with the correlation in absolute terms. Figure 13 depicts both shares in a barplot for graphical comparison.

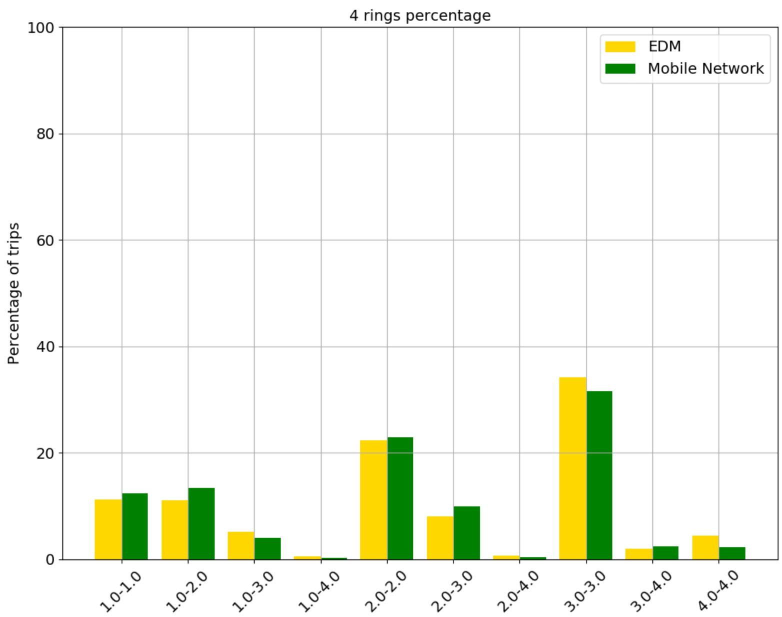


Figure 13. Comparison of the shares of trips for both OD matrices in terms of distribution rings (1 is city centre and 4 is regional area).

3.1.2.6. Trip purposes

Figure 14 depicts the shares of trip purpose combinations (origin and destination) for both OD matrix estimation approaches.

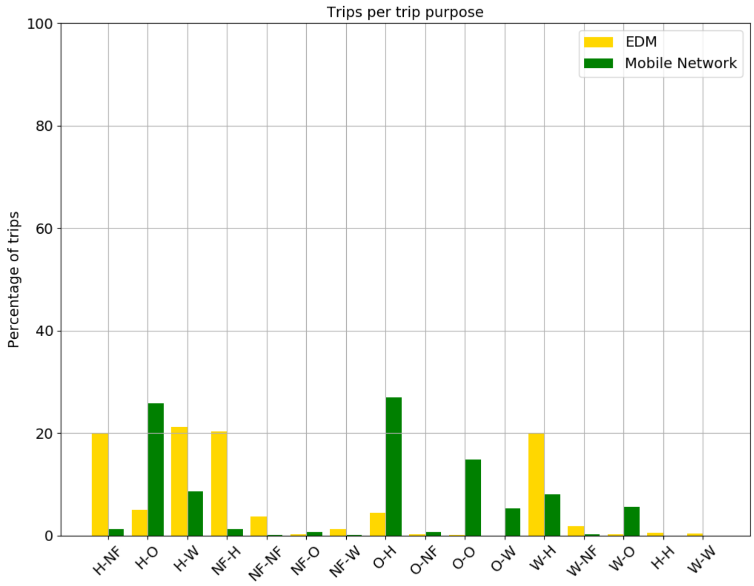


Figure 14. Shares of trip purpose combinations (origin-destination) for both OD matrix approaches.

It is clear that the comparison of O and NF purposes is very sensitive to the purpose correspondence between both sources, which is highly uncertain. However, some interesting aspects can be observed in relation to home-work flows. The survey reports much more H-W and W-H trips than mobile network data. This can be related at least to the following aspects:

- Work trip purpose identification from mobile network data depends on the hypothesis used for selecting certain frequent locations as work candidates for each user. Common criteria include aspects such as time length, frequency in weekdays or starting time of the activity. It is challenging to come up with a set of criteria that covers all the labour situations, e.g., part-time shifts. This may lead to underestimating work activities in mobile network data analysis.
- Trip chaining behaviours are also difficult to measure. Trip chaining refers to the inclusion of certain intermediate activities (e.g., shopping) in between home-work and work-home trips (Primerano et al., 2008). Surveys are known to under-represent 'trip chaining' behaviours (Bricka & Bhat, 2006), either because questionnaires are unclear on the definition of trips or because participants tend to simplify their activity-travel diaries, eluding such intermediate activities. This may lead to overestimating 'direct' home-work and work-home trips in surveys. The proportion of non-home-based trips in the survey is very small, which indicates that the results may be affected by this bias.

3.1.2.7. Temporal distribution

Figure 15 depicts the temporal profiles of both OD matrix approaches at the level of hours of the day.

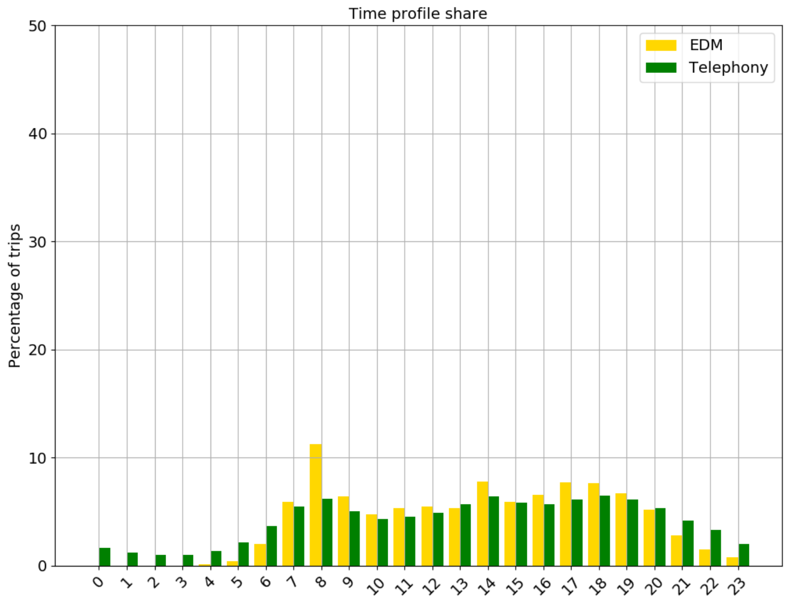


Figure 15. Hourly distributions of both OD matrix estimation approaches.

Figure 15 shows a high degree of similarity between both matrices. In any case, it is worth to mention the much higher morning peak at 8 observed in the EDM matrix whilst the mobile network data one gets fewer trips at that time, and more during the night hours before the peak. Indeed, early morning trips (0-5) in the mobile network data-based matrix do not appear in the survey based one. This could be further explored with other data sources such as traffic counts or public transport smart card data.

It is possible to further segment hourly distributions by age and gender, as shown in Figure 16 and Figure 17.

The main difference is observed for both young (0-19) and elderly (65+) age groups. The former group shows differences through morning hours, indicating in the survey there are almost no movements while mobile phone data indicates a certain range of movements. For the case of the elderly, the profile is in both cases plainer, even though the survey approach is capturing a marked peak of trips in the middle of the morning (9-12) that is not observed in the patterns extracted by mobile network data.

Regarding gender, Figure 17 displays few differences both in terms of the profiles and the comparison between survey and mobile network data-based approaches. The survey reflects a higher proportion of trips in the off-peak hours during the morning for females, a trend that is not captured by the mobile network data. This may indicate some inconsistencies between the actual gender of the device user and the one reported in the contract.

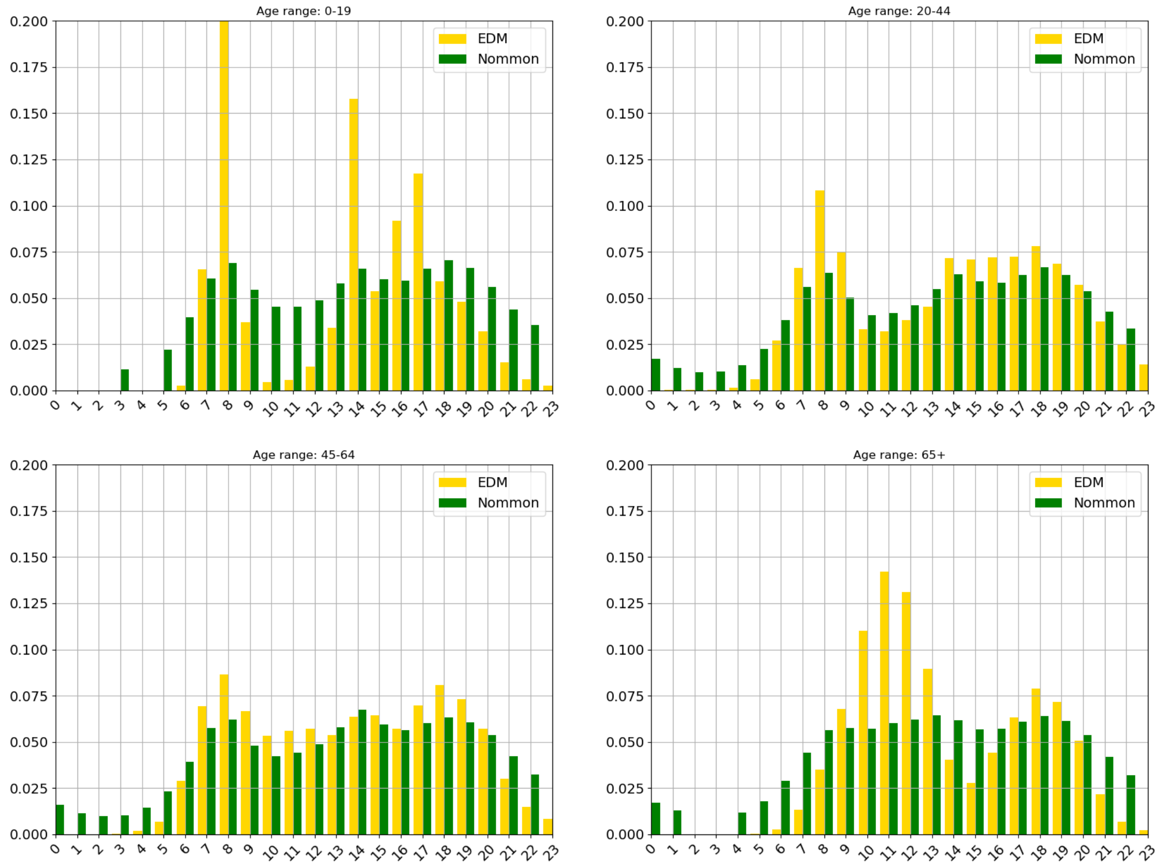


Figure 16. Hourly distributions by age group

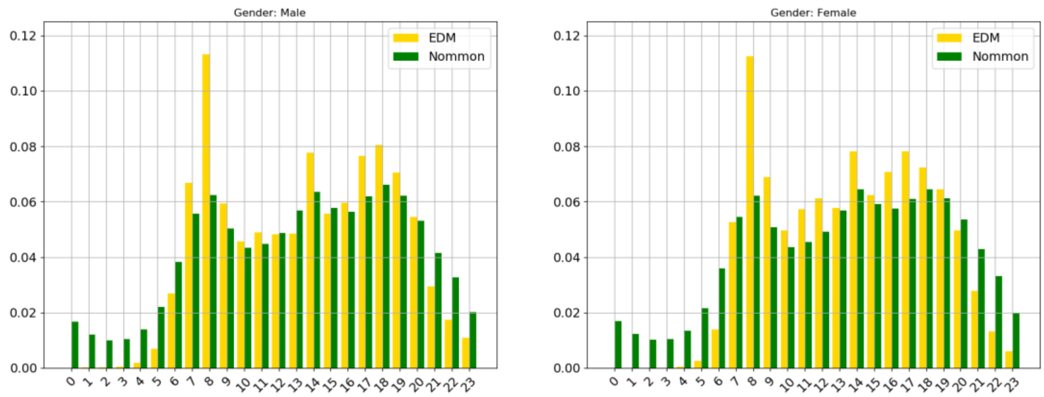


Figure 17. Hourly distributions by gender

3.1.2.8. Conclusions

In this section, two travel demand data sources for estimating OD matrices are compared, one based on a mobility household survey and one based on mobile network data. The comparison of both approaches has been performed over different features of the matrices, such as the volumes of trips, hourly distributions or distances as well as additional segmentations based on sociodemographic and trip characteristics like trip purpose or gender or age.

In general, both OD matrices have shown fairly high degrees of similarity in most dimensions, suggesting that both alternatives provide fairly good estimations of the region OD matrix. In any case, the analysis has shown some discrepancies in certain segmentations, mainly trip purposes and part of the hourly distribution. These aspects deserve further research for the improvement of the algorithms employed to extract OD matrices from mobile network data. In particular, the following elements can be highlighted:

- Age and gender segmentation of the mobile network data is limited by the fact that the profile of the user of each device may not always correspond to the age and gender reflected in the contract information. This can be corrected by looking at the characteristic mobility patterns of each age range and gender in the survey data that can be identified in the activity-travel diaries, in order to relabel agents accordingly.
- The number of trip purposes that can be identified through mobile network data is limited. This is mostly based on the frequency of visits to certain locations. The use of land use data and points of interest may be useful for complementing this methodology. More research is needed on the identification of work purpose from mobile network data, particularly with regard to part-time shifts.
- Some coverage areas in rural areas may be too large to identify intrazonal trips. The spread of 5G technology is likely to increase the number of antennas, which in turns reduces the average size of the coverage areas.

3.1.3. Shared-mobility adoption

3.1.3.1. Motivation

Shared mobility services have suddenly appeared in the last years in many cities, and the profiles of those people adopting them have been explored in the literature. Nevertheless, the data sources available for the case study provide the perfect scenario for exploring those insights for the Madrid region, which has been one of the cities with the most activity of shared mobility services.

This section explores survey data from the Madrid mobility survey as well as user diaries extracted from operation data for BiciMAD and Muving services following the methodology developed in Section 2.4. With these data sources, shared mobility adoption profiles are explored in the survey data while user income distribution inside and outside services geofences are explored with the operation data. In addition, the results for the home inference methodology proposed in Section 2.5 are provided.

3.1.3.2. Shared mobility adoption patterns from household survey data

The 2018 regional mobility household survey already includes some trips made in the existing shared mobility services, which provides a sample of users of these emerging mobility solutions. In total, the survey collects 16,541.69 expanded trips (223 sample trips in the survey) in shared mobility performed by 10,959 expanded travellers (143 survey respondents). This accounts for a 0.16% of the total survey respondents.

3.1.3.2.1. Age and gender

The age distribution of shared mobility users and the entire population of the region is shown in Figure 18. The figure clearly shows how shared mobility users belong predominantly to the age group 20-44, being the second largest the 45-64. As expected, shared mobility users in the youngest group (0-19) are scarce, possibly due to the barriers for the use of these services (e.g., age limitations, driving license required) as well as the cost of services being higher than alternatives such as public transport. Regarding the elderly age group (65+), the share of the group is also significantly lower, possible due to a reluctance towards the adoption of new technologies and the effect of the digital divide that has been previously noted in the literature. As a result, middle aged groups are overrepresented for shared modes with respect to the general population distribution whereas young and elderly age groups are much underrepresented in shared modes.

Regarding gender, the number of male users is 61.1% of the total against 38.9% female users. In contrast, the share of females in the overall survey population is 52.2%, which makes the male-bias very high. This is consistent with the literature, where it has been extensively shown that more shared mobility service users are males, like in (Becker, Ciari, & Axhausen, 2017) in Basel or (Yoon, Cherry, & Jones, 2017) in Beijing.

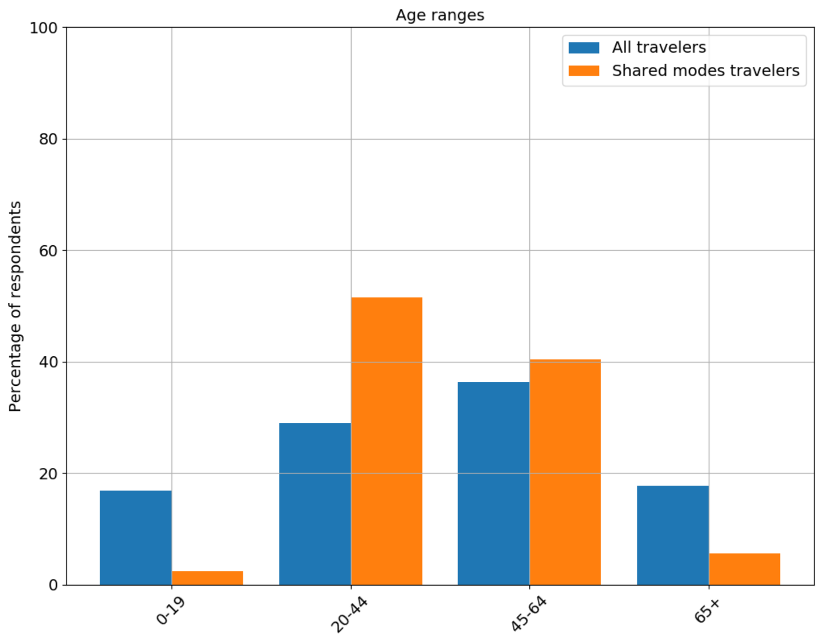


Figure 18. Comparison of travellers share by age range: all travellers vs shared mobility travellers

3.1.3.2.2. Education and labour status

Education level is another important parameter from shared mobility users that has been observed in the past. Figure 19 displays the share of each level of study compared between all travellers population and shared modes users.

The figure points to most shared mobility users having university degrees, surpassing the general population by approximately 35 basic points, whereas in the rest of study levels the share of shared users is generally smaller or comparable to the general population share. This observation is in line with most of the literature that states shared mobility users tend to have higher education (Kortum & Machemehl 2012).

Finally, the professional activity of each travellers is reported. There are three questions in the survey dedicated to the professional details of each respondent: dedication, activity and area. Dedication relates to the professional state of the person (working, seeking, retired), activity denotes the nature of work each person develops (public or private sector, entrepreneur…) and sector provides the sector from a reduced list. Figure 20 depicts the share for the different categories of each of these dimensions in the shared user population.

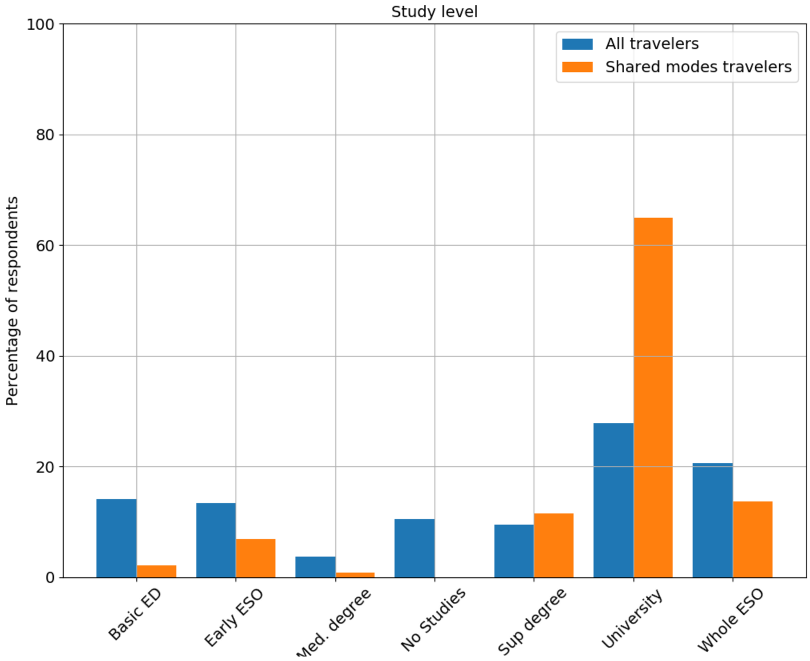


Figure 19. Comparison of the frequency of each study level for all population and shared modes users

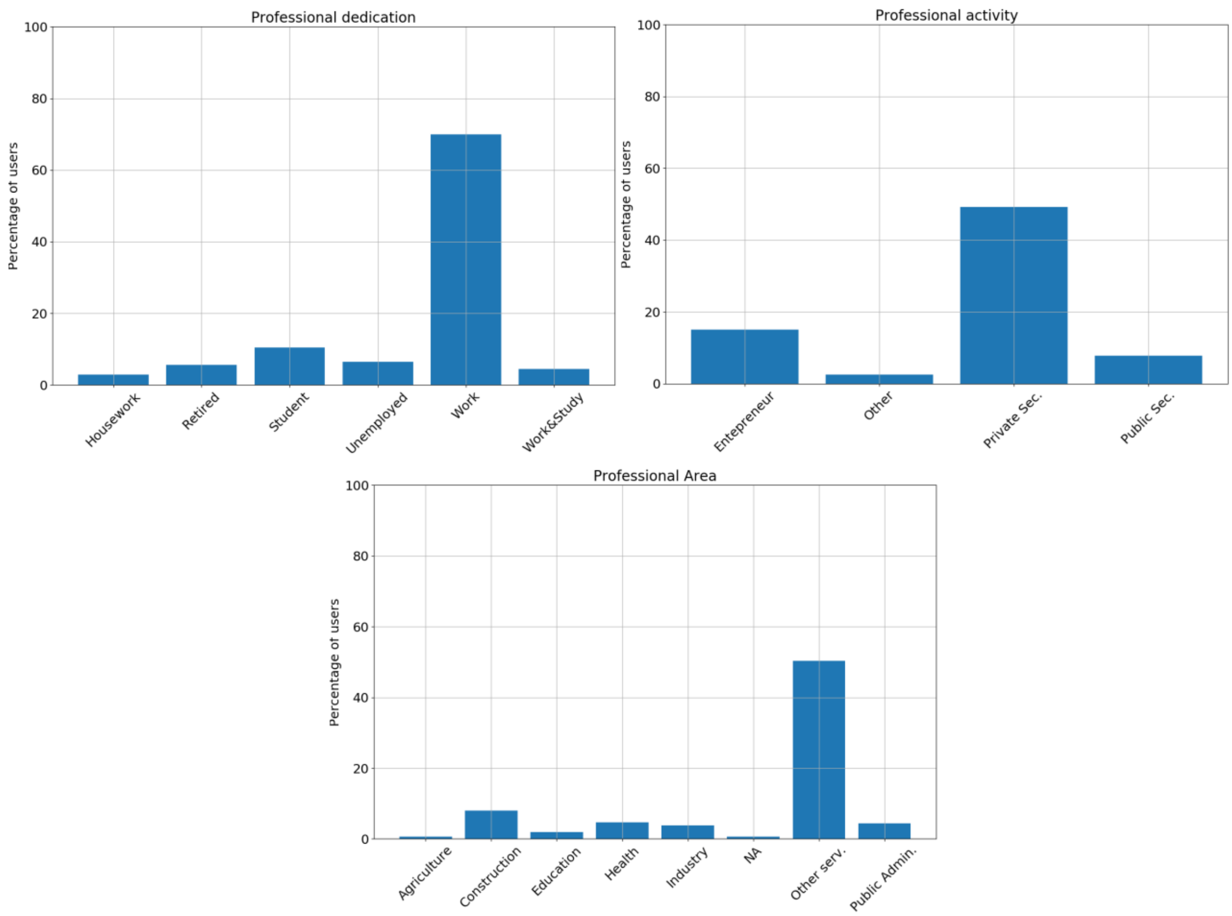


Figure 20. Professional dedication (top left), activity (top right) and sector (bottom centre) shares of the shared mobility users

In general, most of the users are working (70%) with a small share of students (approximately 10%), possibly due to the additional cost that shared mobility services entail and the age limitations for using the services. Regarding activity, most users are dedicated to the private sector, even though there is a significant share of entrepreneurs, typically involved and participating in new technology-based services and thus matching the profile of early adopters. Regarding area, the most relevant area is that of services, which employs 50% of shared mobility users. Actually, this area appears to be the most suitable for technology-related jobs, which will be occupied by people that is, again, more familiar and potentially more interested in new services.

These figures clearly indicate a user profile prone to technology, possibly involved in new technology-related businesses and working on private companies, maybe even their own. For reference, Figure 21 compares the professional dedication structures of all the travellers and the shared mobility users.

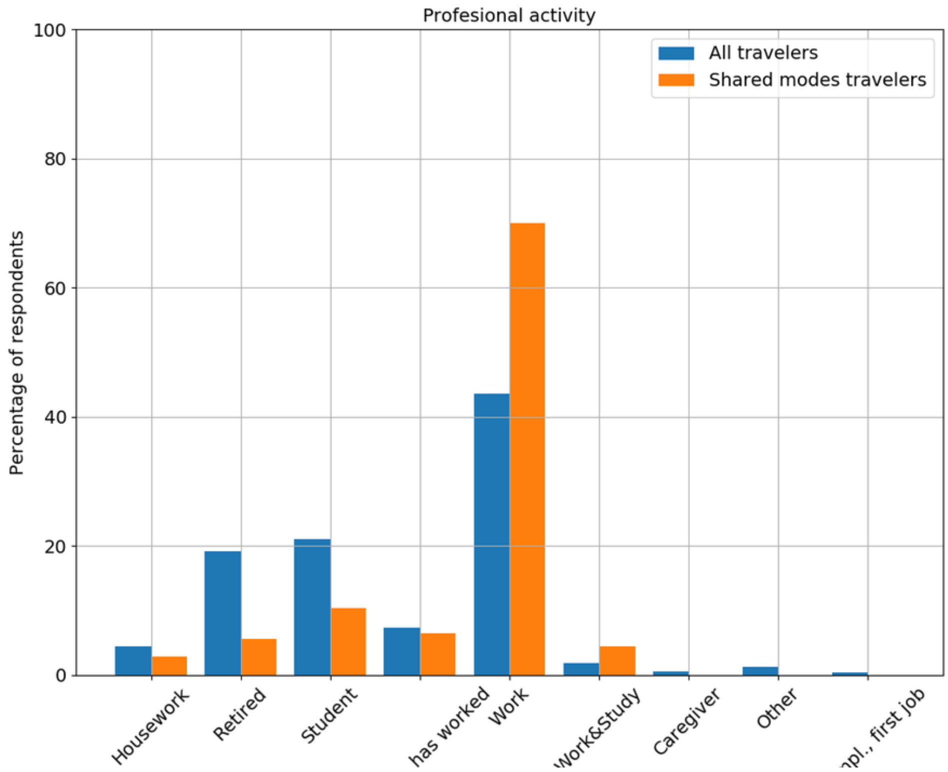


Figure 21. Comparison of professional activity shares between shared mobility users and all region travellers

Indeed, the figure shows that shared mobility users are working in larger proportion than all the travellers. Apart from that, the difference between the number of retirees in the general population and in shared users is relevant, indicating a clear underutilisation of shared modes by the elderly. The same underutilization applies for the students.

3.1.3.2.3. Public transport card

In Madrid region there are two types of public transport card: the TTP and the MULTI card. The TTP is a personal card that can be loaded with the regular monthly pass ('abono') that enables the use of public transport with no usage limit. The MULTI card is a non-personal card that can be loaded with different tickets, mainly single-trip and 10-trip tickets, each one providing one-time access to any public transport service. Figure 22 displays the comparison of the share of people that owns one, both or None of the public transport cards.

The figure displays that many shared mobility users declare having some kind of public transport card, being the MULTI more frequent than the TTP in their case. The reader should note that TTP card ownership entitles their owners to special fares to the annual subscription to the BiciMAD bike sharing service.

Comparing modes, we can observe differences in the predominant type of public transport cards, being TTP the most frequent for all travellers and MULTI the one for shared mobility users. Furthermore, the difference between people that declare not having either card is significant, being the frequency of shared mode users much lower. This suggests most shared modes users do consider public transport for some of their trips as they own some card for its use.

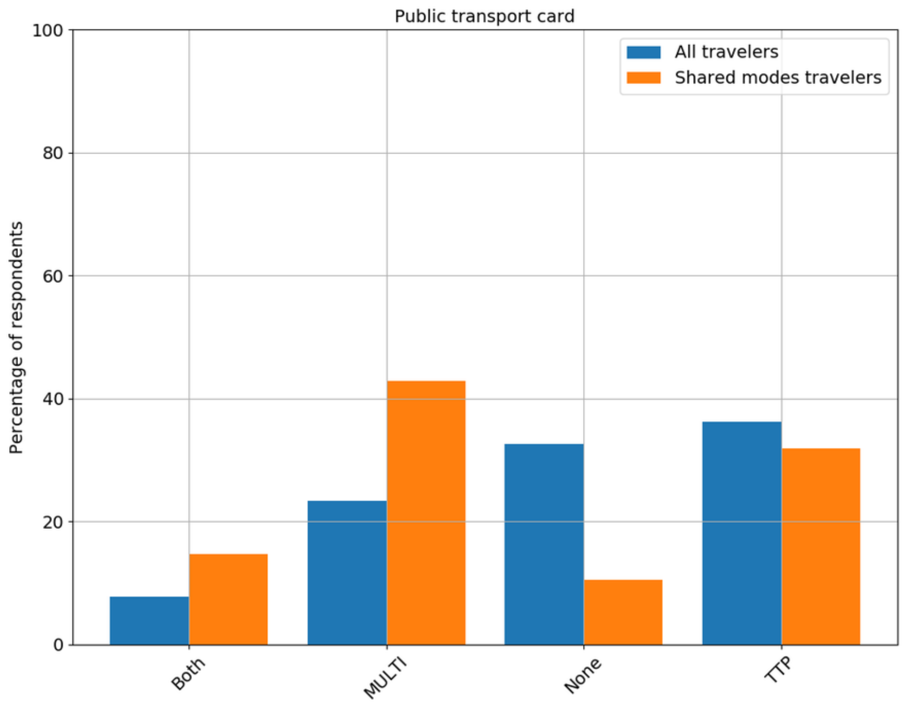


Figure 22. Public transport card ownership share comparison between shared mobility users and the entire population

Regarding the penetration of the monthly subscription ('abono') among shared mobility users, 51.5% of the responses are missing in the survey, probably because those users lack the proper TTP card. From those who have the card the number of users with a monthly subscription is 38.3% of the total vs. only 10.2% of users not having it. With respect to modes, we can observe in Figure 23 below the user card ownership frequency by shared modes compared to any mode users.

It can be observed that most bike sharing users own some kind of public transport card, with a penetration of the TTP equal to that of the general population and the MULTI card being the most common far above the general population. In fact, bike sharing users are the group most represented in all categories that have at least a public transport card. The larger share of TTP holders can be related to the existing discounts for the users of this card in the bike sharing system.

Another relevant observation is that car sharing users consider public transport more than the general population as well: their share of people with no public transport card is lower than that of the general public, being the MULTI the predominant card with a larger difference with respect to TTP than in the bike sharing case.

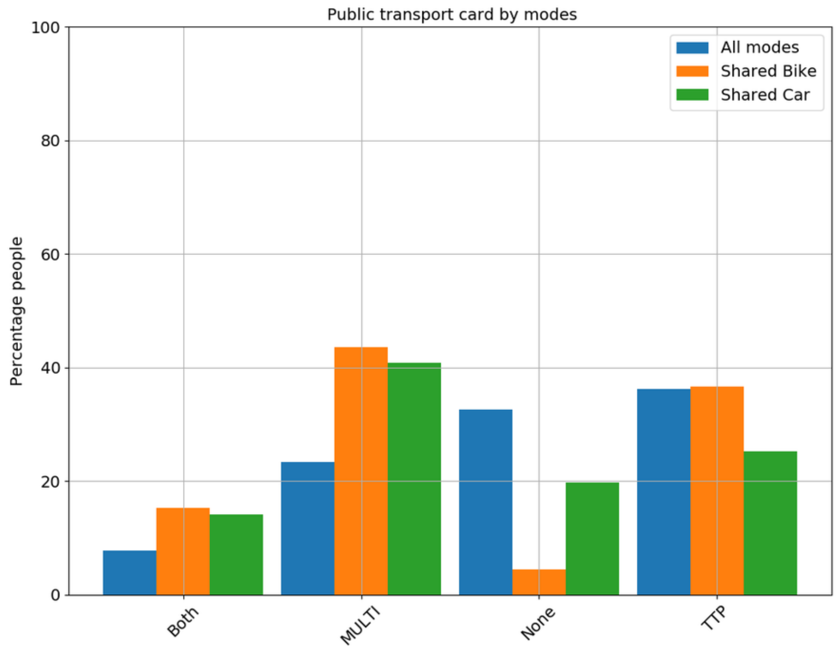


Figure 23. Public transport card ownership by relevant shared modes

3.1.3.2.4. Driving license

Survey respondents were asked whether they have any kind of driving license, of which they can answer: "no license", "little scooter" (A1/A2 driving license types), "motorbike" (A driving license type), "car or larger vehicles" (B, C, D, E driving license types) and car and scooters (A and B driving license types). This question is relevant since a license is required for cars and motorcycles in Spain. Figure 24 compares the share of people with each type of license between the entire traveller population and the shared modes travellers.

As observed, most users have predominantly car driving licenses (B, C, D), the most useful license type for all shared modes since after having it for two years, its holder automatically acquires the license to drive small motorbikes like the ones used by shared mobility services.

In general, a larger amount of car driving license holders is observed among the shared mobility users and also a significantly smaller proportion of users that have no driving license at all. For further details, Figure 25 provides the bar plot of the share of license holders by shared mode.

The figure further shows the prevalence of the car driving license for both shared modes (bike and car sharing). The reader should note that travellers underage (less than 18 years old) that cannot obtain a driving license have been filtered out to normalize the conditions on the comparison. Even in the case of bike sharing, the users with no driving license are proportionally less than the group all mode users.

Actually, while car sharing users do have the car driving license, bike sharing users present the same amount of car driving license holders than all mode users and larger penetration of motorcycle driving license holders (A type), being the number of users with no license smaller than the general population.

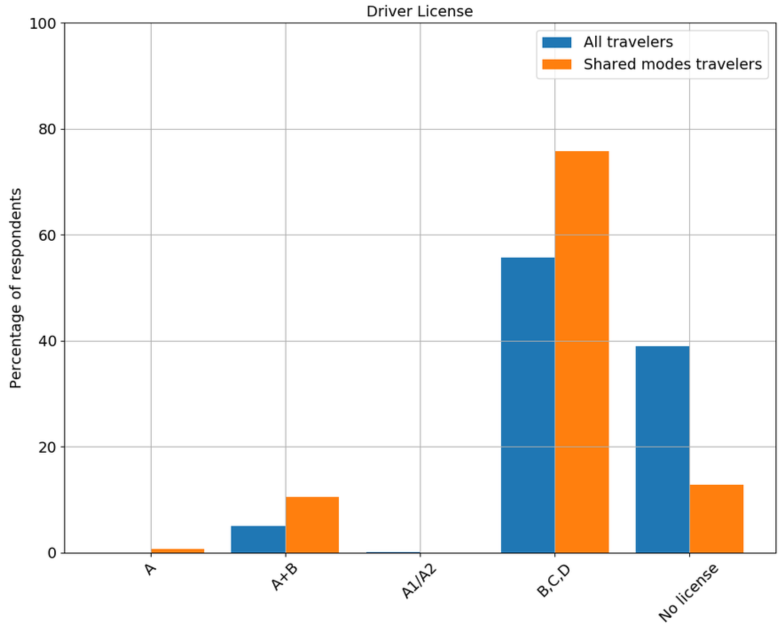


Figure 24. Comparison between shares for all population and shared mode users

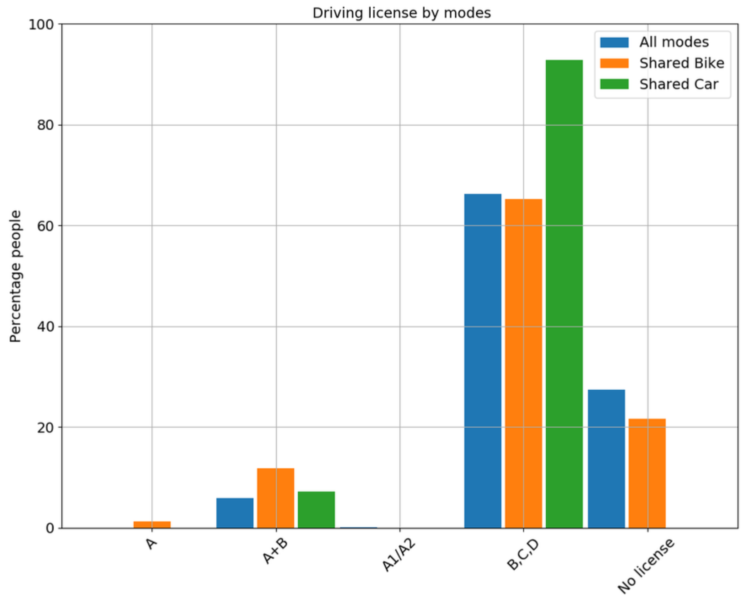


Figure 25. Share of type of driving licenses ownership by shared modes

3.1.3.2.5. Conclusions

In general terms, shared mobility users in Madrid according to the most recent mobility household survey match the user profiles observed in the literature, since most of them are 20-64-year-old males with higher education, mostly university degrees. In general, shared mobility users work in the private sector in areas potentially associated to services, possibly with a high technological component, which makes them potential candidates to be early adopters of disrupting new services like shared mobility. Most shared mobility users have some type of driving license, but interestingly, there is a large proportion of shared mobility users that own some type of public transport card, especially in the bike sharing segment.

3.1.3.3. Validation of shared mobility users home inference methods

User data from two shared mobility services is available for the Madrid case study. Concisely, BiciMAD (bike sharing) and Muving (moto sharing) services have provided data for analysis. In both cases, user data includes the home location at the zipcode level, so the home inference methodologies proposed can be validated. BiciMAD presents the home location information attached to each trip, and the user identifier is re-anonymised every day. This implies that only cross-sectional methods (Section 2.5.2.1) can be validated. Muving presents the home location information together with a customer database whose user identifiers are consistent with the trip data, thus both cross-sectional and longitudinal methods (Section 2.5.2.2) are applicable.

Nevertheless, there are some restrictions in the user home locations provided that constrain the results as follow:

- A considerable number of users (54.8% for BiciMAD and 40% for Muving) do not report user location and, therefore, cannot be validated.
- Not all the users with home location live inside services' geofences, so no home location will be able to correctly assign home to those users living outside the geofences.
- Some reported home location zipcodes are wrong (more than 5 digits, include letters, etc.) and have to be dropped from the analysis.

Hence, the performance metrics obtained for home location inference at the zipcode level will have been filtered beforehand to ensure that only those elements with a presumably correct "ground truth" value have been used for validation.

3.1.3.3.1. Results from home inference based on morning hour range

The first proposed methodology for home location inference consists on the assignment of home location to the origin of the first trip that occurs within the range 6 AM - 10 AM in the morning, which would be the trip from home to work or study. The number of users with trips within this range is 24% for BiciMAD and 22% for Muving, so this methodology is constrained by the fact that it is only applicable to a limited number of users. Table 5 depicts the obtained accuracy for this method.

Service	Total users validated	Accuracy from users with home location
BiciMAD	67,102*	51.7%
Muving	4,937	45.7%

Table 5. Summary of the results for the morning hour-range based home inference

* BiciMAD reported users are "user-days" since longitudinal analysis is not feasible due to daily re-anonymisation.

The table clearly shows that the accuracy of this methodology is low. This approach gets the declared home location for around half the users, which implies one of each two predictions of the system would be wrong.

3.1.3.3.2. Results from home inference based on roundtrips

The second proposed methodology is based on roundtrips, that is, pairs of trips where the first trip origin matches the last trip destination. Concisely, most frequent roundtrip location within the user-day is assigned as home location. If no roundtrips are found, the previous morning range criteria is applied if possible. Table 6 summarises the results of this approach for each of the services under study.

Service	Total users validated	Accuracy from users with home location
BiciMAD	126,620*	52.4%
Muving	5,562	51.5%

Table 6. Summary of results for roundtrip based home inference

* BiciMAD reported users are "user-days" since longitudinal analysis is not feasible due to daily re-anonymisation.

The table clearly shows a slight improvement using roundtrips instead of morning-range trips, but still the home inference abilities are much limited to half of the total users, which is not good enough to correctly perform user characterisation correctly.

3.1.3.3.3. Results from location recurrence based home inference

Location recurrence is based on selecting the most frequent location from longitudinal data within the dates of study, focusing especially on the most frequently repeated location, the most frequently repeated origin location within the morning range and the most frequently connected (origins or destinations) location. Table 7 summarises the resulting accuracy of selecting one of the two most frequent locations as home location for each of the cases.

Case	Option 1 accuracy	Option 2 accuracy
Location recurrence	14.8%	31.8%
Location recurrence in the morning range	11.4%	6.2%
Connected locations	9.7%	21.6%

Table 7. Accuracy of the two most frequent recurrent locations for each category

Surprisingly, the table indicates that the approaches of location recurrence and connected locations work better with the second most frequent location rather than the first one. This could be happening due to location adjacency: the most convenient vehicle for users is more often available within a location adjacent to where they live, so the most frequent option does not match with their home location, but an adjacent location.

In any case, even the cumulative accuracy from two first options are not good enough to provide the consistent home location inference required for the problem.

3.1.3.3.4. Machine learning based rule inference for home location

The weak results obtained by the heuristic methods explored above suggests that the systematic evaluation of all the potential variables for inferring home location (e.g., morning range trips, location frequency, etc.) is needed. This has been done through decision tree models showing to what extent each variable may be related to the home location information. A dataset has been generated according to the guidelines in Section 2.5 for both shared mobility sources and a decision tree model to determine whether an observed zipcode is the home location of a user has been trained and its decision path plotted. In both cases, the training process has been performed over the users that performed trips during the periods of the months of study (October and December 2018, February, April, June, July, October and December 2019 and February 2020). To provide the best possible model, hyper parameter tuning has been performed to maximise the training F-score of the model (the F-score is a measure of a test's accuracy and it is calculated from the precision and recall of the test. The highest possible value of an F-score is 1, indicating perfect precision and recall, and the lowest possible value is 0, if either the precision or the recall is zero).

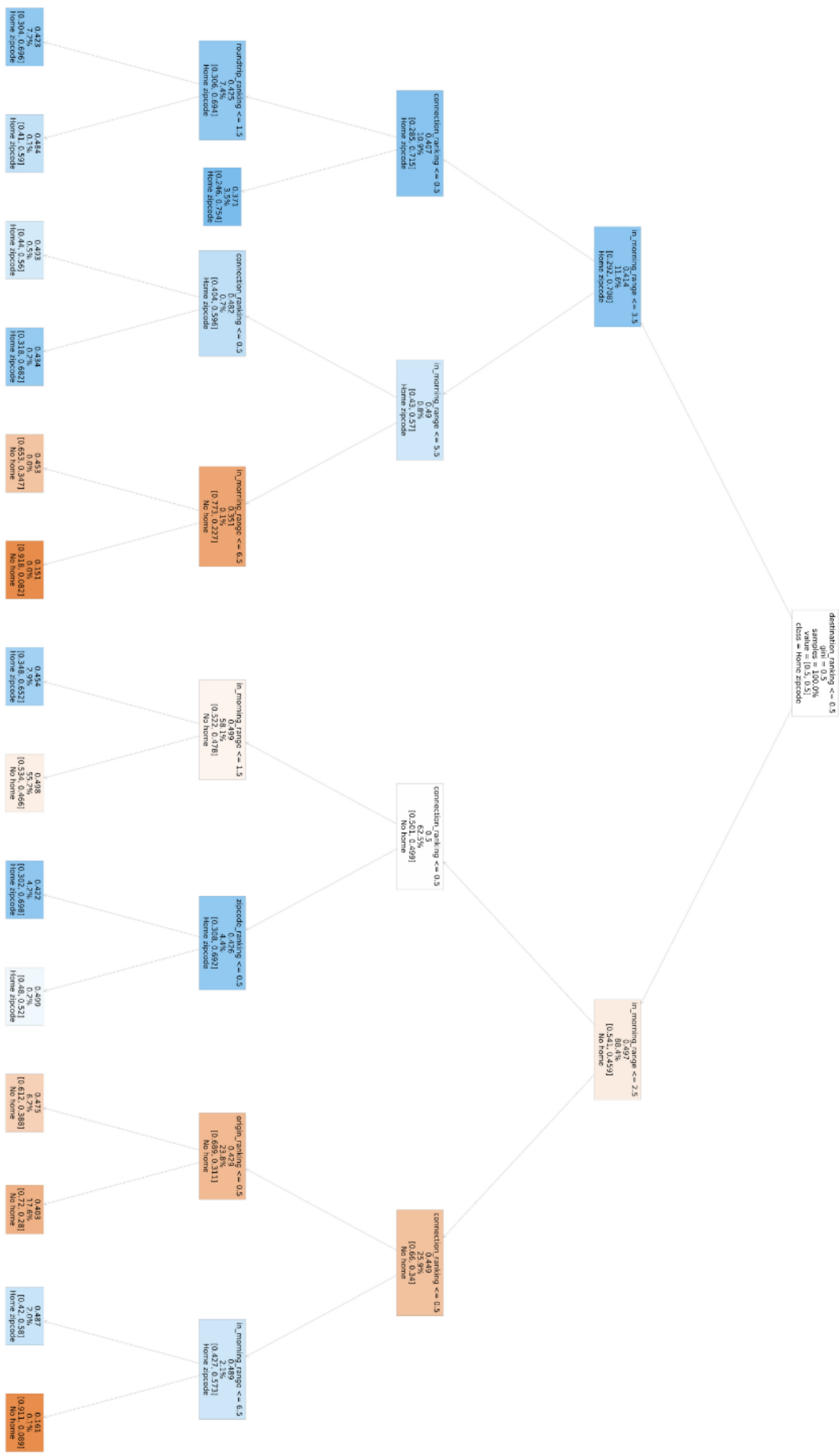


Figure 26. Inferred rule hierarchy based on ranking features

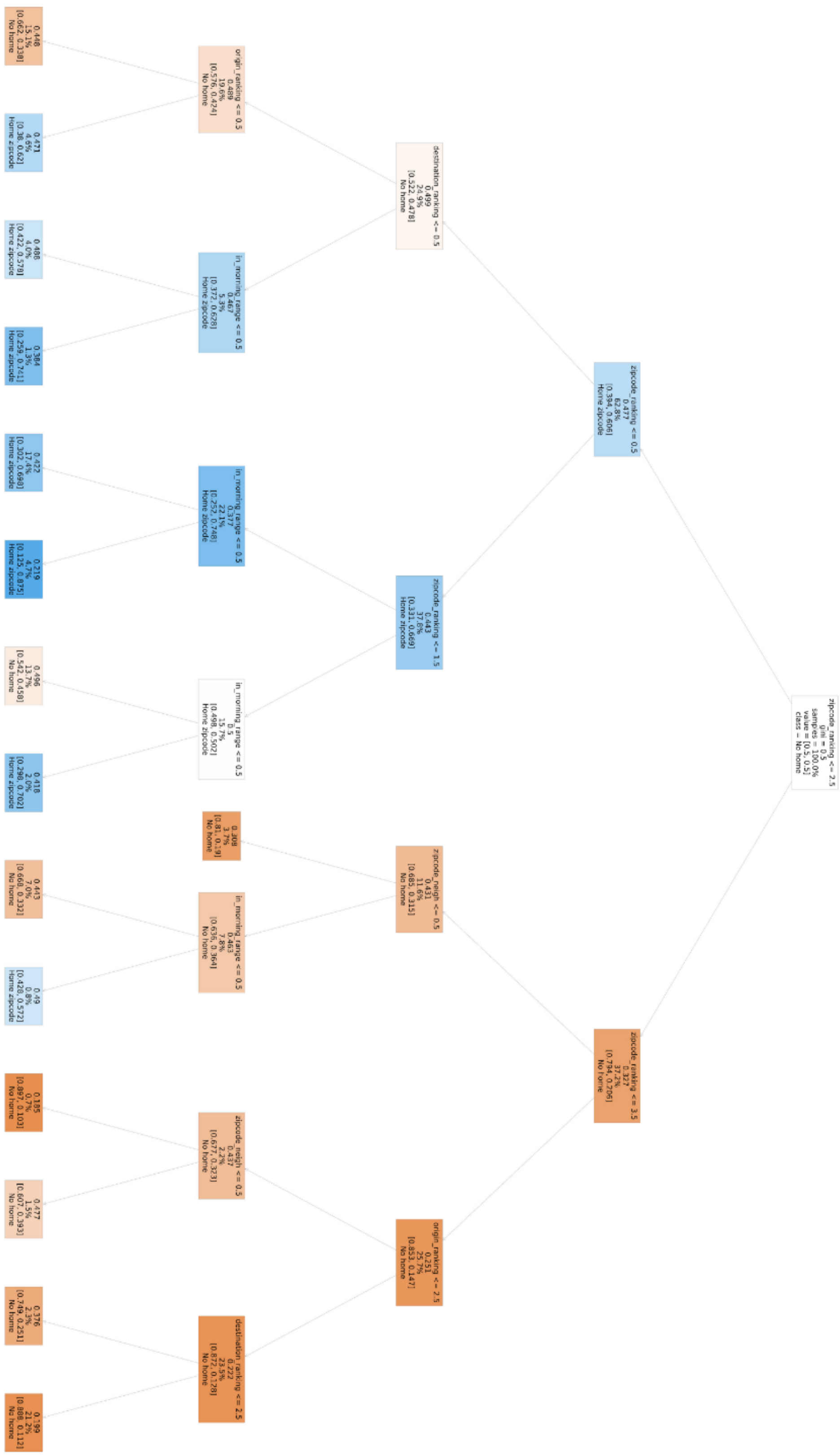


Figure 27. Hierarchy of rules for the model trained using Muving data

Even if BiciMAD only has cross-sectional daily data, the approach has also been tried to see if any pattern related to non-longitudinal variables can be extracted from each ‘user-day’. The resulting tree accuracy and F-score is 0.32. Figure 26 depicts the learned tree structure for the BiciMAD case.

The resulting tree, as shown in Figure 26, is able to perform separation based mainly in the position within the destination ranking, especially for zipcodes in the initial position, and whether the trip occurs during the morning range or not. Other interesting rules that appear more times are being first in the ranking of most connected zipcodes and, less frequently, being first in the rankings of origin and destination zipcodes.

Regarding the decisions performed by the model, the leaf node containing the greatest number of points is also the most uncertain (samples from both classes are nearly 50-50%). Still, there some leaf nodes that yield a collection of points with clear majority of the class selected by the algorithm, like the sub-tree following the rules below first position in the destination ranking, not in the morning range and first in the connection ranking (the zipcode most connected to others), which aggregates up to 24% of the points, is classified as no home and has a proportion of 70% no home, 30% home zipcode.

Muving rules for home inference

For the Muving case, the trained model yields an accuracy and F-score value of 0.42, which supports slightly better the rules inferred. Figure 27 plots the tree-based rule structure followed by the training classifier.

Concisely, it can be observed that the most relevant variable is the position in the ranking of the most frequent zipcodes along with whether a zipcode appears in the morning range and the position of each zipcode within the ranking of most frequent roundtrip locations. Origin and destination rankings have some relevance as well. The best neighbouring positions in rankings are seldom considered as they appear fairly down in the tree.

In general, it is possible to observe that the major separation is given by the initial comparison, whether the observed zipcode is located among the top three positions of the general zipcode ranking. From there, most rules on the left tree end in a zipcode which corresponds to a home location whereas the rules of the trips on the right are dominated by zipcodes not being home locations. Despite, it is worth noting the leftmost path on the tree, which identifies a relevant “no home” position of 15% of the points from the dataset, even though class balances are not very declined to either class. The points assigned to this position are those between the three top positions in the general ranking and top positions in both the destination and origin ranking.

3.1.3.3.5. Conclusions

Home inference experiments have shown that the variables that can be extracted from shared mobility operation data are not suitable for consistently identify user home location with an acceptable degree of accuracy. Different alternatives based on user trip locations have been tested, being none of them good enough to provide accurate home inference results for a given set of users with their home location.

Machine learning based rule inference has been tested as well, showing most of the variables considered unable to clearly differentiate between regular origin or destination zipcodes and users’ home locations. In any case, some observations can be derived from all the experiments:

- Home location are typically among the most recurrent zipcodes a user visits, being most user homes located within the three most frequently repeated zipcodes in the Muving service. This can serve to discard other locations.
- Longitudinal trip information provides much more information than the trips on a single given day. Users in general perform few trips a day, whereas having the trips for more than one day facilitates the unveiling of frequently visited locations.

- It is clear that not many users take shared modes on their regular trips (home-work), as recurrent use patterns are not very frequent. There is up to 20-22% users from both services that perform trips during the peak morning range and up to 42% (BiciMAD) or 33% (Muving) users that take roundtrips in the services through days. The use of these services for other trip purposes rather than home work is proven to be a challenge for home location inference methods.

3.1.3.4. Shared mobility users income profiling

3.1.3.4.1. Income distribution by services

Some authors have highlighted that shared mobility users are typically high-income workers, in the line with the observations from (R. Clewlow, 2016), (Becker, Ciari, & Axhausen, 2017) and (Yoon, Cherry, & Jones, 2017), probably as a result of city centres relevance for service deployment and the ability of these individuals to have money to spend in this type of services. The analysis of the household survey as can be seen in Section 3.1.3.2.2 already suggests that the profile of users is biased towards workers in the case of Madrid, but the household survey does not include income information. Therefore, the methodology explained in Section 2.5.3 will be used to assess the income distribution of users. Since BiciMAD does not provide a customer database with home locations, but this information is instead attached to each trip, the method explained in Section 2.5.3.2 is the one applicable. For Muving, there is this database, so both methods (customer database based and trip data based) can be used.

Two pre-processing operations have to be conducted with the available datasets to develop the analyses:

- **Obtention of income information at zipcode level**: income data for the experiments are obtained from the Spanish “Household Income Distribution Atlas”. This information is provided at a census tract level. Since home location for both BiciMAD and Muving is available at the level of zipcode the income data was treated to match this level of aggregation. Census tracts were geographically intersected with zipcodes and the average income average of the Census tracts contained inside each zipcode was assigned to it. Since Census tracks are not fully contained in zipcodes the amount of area of each tract that is contained within the zipcode was used as a weight to perform the average. This average income was directly mapped from each user home location zipcode into a figure of income. Finally, to constrain the analysis to the city of Madrid, zipcodes from outside the city of Madrid were filtered out.
- **Extraction of shared mobility service area or geofence**: The geofence is the area or groups on areas where the shared service operates. For free-floating services, the geofence determines where vehicles can be taken or left at any time. In the case of station-based services, the geofence is not needed, as the only places where vehicles can be taken or left are stations. However, for analysis purposes, this geofence can be artificially generated. Once the geofence is known, it can be intersected with the zipcodes to identify which users of the service live inside it and which ones outside the service area. In the case of Muving the geofence is available in the service webpage as a shapefile. On the contrary, BiciMAD geofence had to be derived. This was done by computing the intersection between the service station locations and the Madrid zipcodes (Figure 28 and Figure 29 show the geofence of BiciMAD and Muving services in the city of Madrid using zipcode areas).

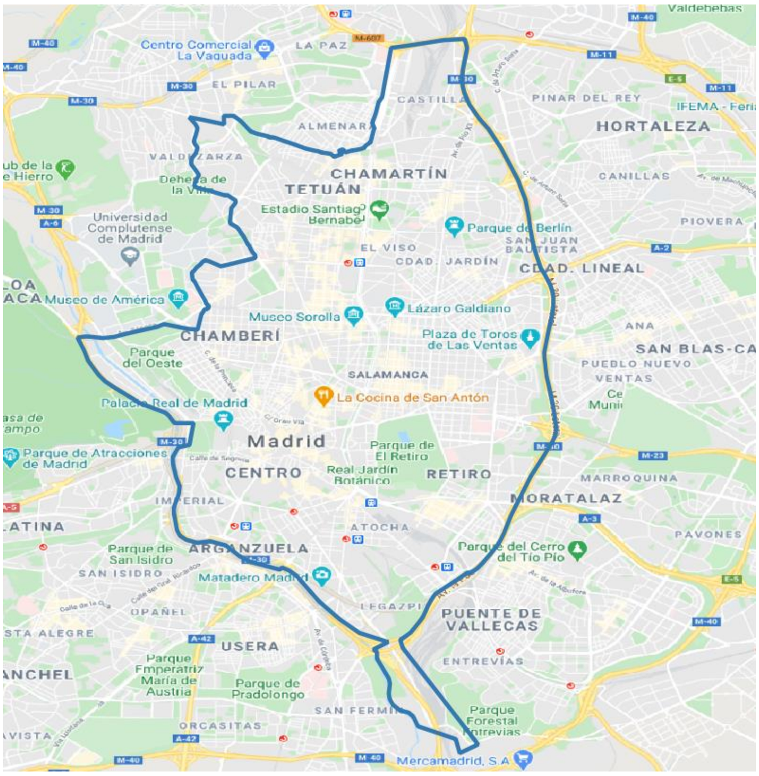


Figure 28. Extracted geofence for the BiciMAD service in Madrid city

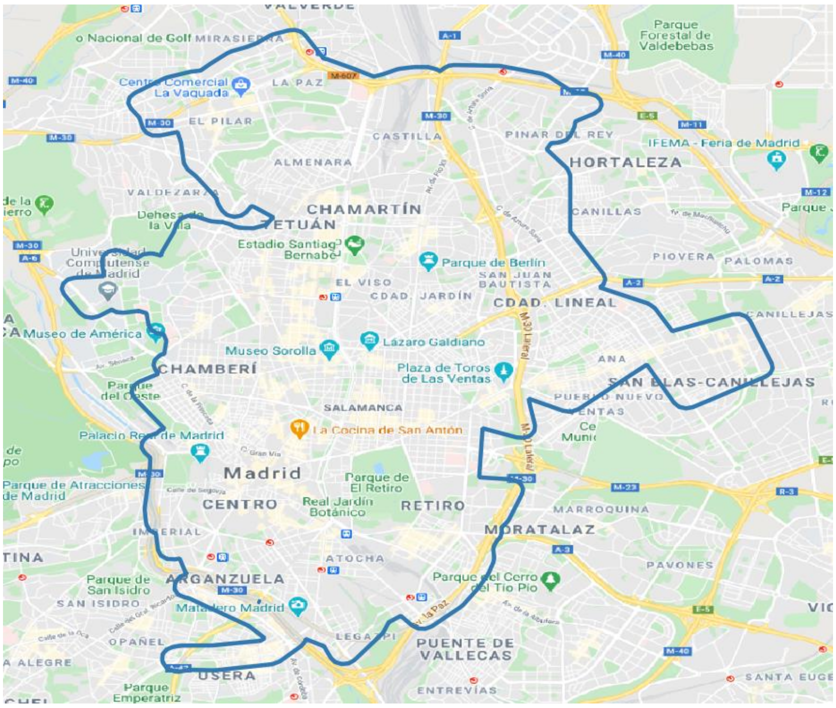


Figure 29. Muving service geofence for the city of Madrid

Once the home location of the users respect to the geofence was identified, user income distribution has been explored in three different setup scenarios: (1) all users together, (2) users living inside the service geofences and (3) users living outside service geofences in the rest of the Madrid municipality. This is done applying the methodology explained in Section 2.5.3.2, since it is the one applicable to all systems regardless if a customer database is available or not (the latter is the case for BiciMAD). The next figures display the boxplots comparing the distributions for BiciMAD (Figure 30) and Muving (Figure 31) services along with the general population distribution for the three cases. Figure 32 presents the boxplots for Muving and BiciMAD users' income distribution for a better comparison.

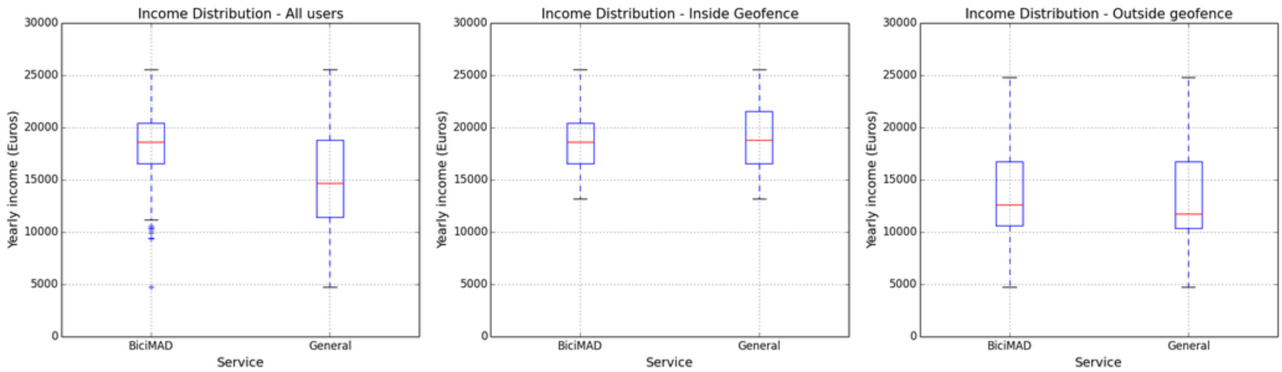


Figure 30. Boxplot distribution for the income of BiciMAD for all service users (left), users living inside geofence (centre) and users living outside geofence (right)

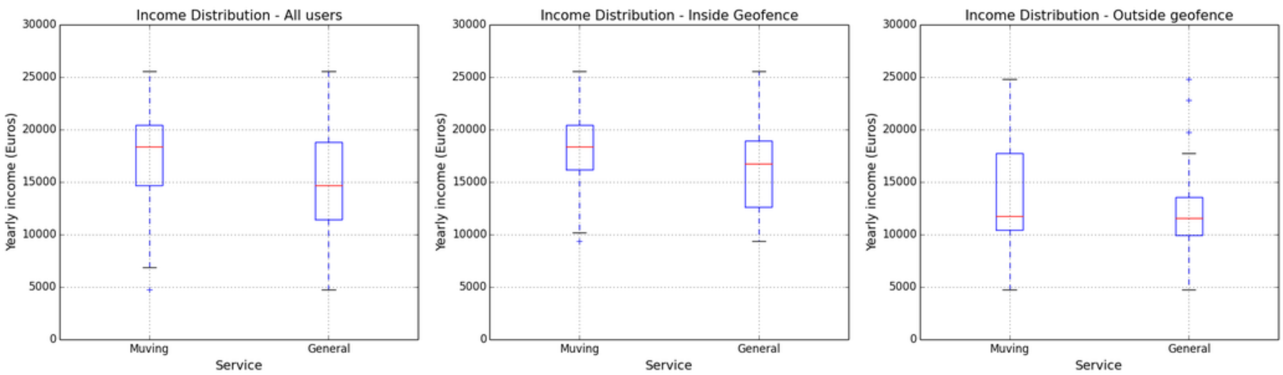


Figure 31. Boxplot distribution for the income of Muving for all service users (left), users living inside geofence (centre) and users living outside geofence (right)

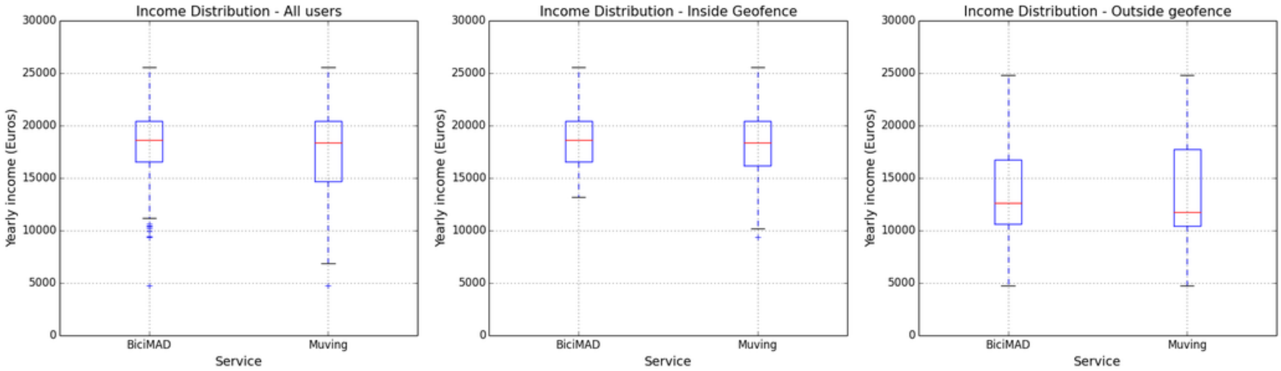


Figure 32. Boxplot distribution comparison for BiciMAD and Muving. The cases are: service users (left), users living inside geofence (centre) and users living outside geofence (right)

The inferred income distribution of the users of both services shows a higher average than the overall population. This is mainly related to the fact that the average income within the service area or geofence is generally higher than outside it, and most users live within the service area.

However, an important difference arises between bike sharing and moto sharing services when looking at those living within the service area. Muving service users have an average income higher than the distribution of the income in its geofence, according to this methodology (9% over). BiciMAD shows the opposite trend, with the average income of users being slightly lower than the population (2.6% below). However, the average income of both services is quite similar as can be seen in the last set of pictures. As it can be seen in Figure 33, the moto sharing service covers areas with less income, but according to these results, the residents in these areas (not served by BiciMAD) have not adopted the service as much as the ones living in higher income areas. This suggests differences in terms of affordability between both services. It will be interesting to see if the expansion of bike sharing services to these districts in recent times follows the same pattern or not.

3.1.3.4.2. Average income and user penetration by zipcode

Despite income distribution suggests there is indeed a difference between inside and outside correlation, it is possible to explore whether shared mobility services increase their user base when the income is higher.

For that purpose, the scatter plot of the average income at each zipcode zone against the share of users captured from the population there could provide more detailed insights individualized by service. Figure 34 provides such a plot for the BiciMAD case and Figure 35 does the same for the Muving case.

In the case of BiciMAD, there is no significant correlation in any of the three cases, which could be indicating that the service has settled among the general public consistently without being influenced by the income at each zipcode zone. At this point, the reader should bear in mind that the BiciMAD service was the first shared mobility service deployed in the city of Madrid and, also is well suited for integration within the public transport system, as many stations are nearby public transport stations. Besides, it has been observed in the Madrid mobility survey (EDM) that bike sharing services (mainly BiciMAD) is highly involved in multi-modal trips including public transport (up to 22% of its trips).

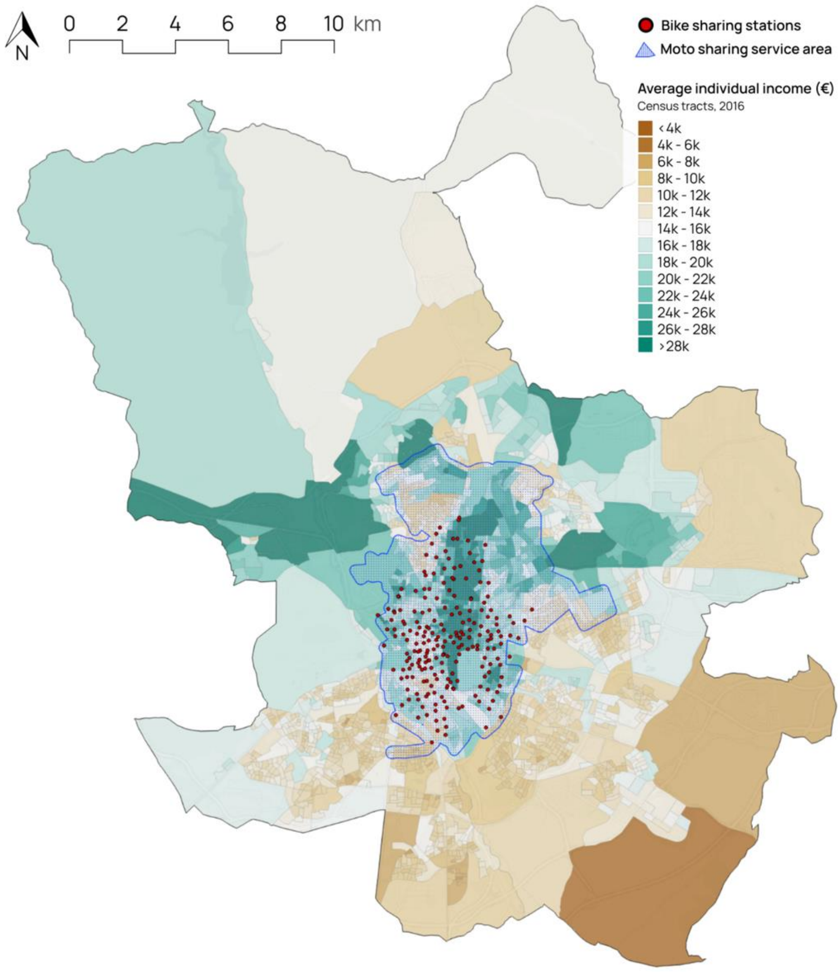


Figure 33. Income distribution at census tract level in Madrid municipality and areas covered by the two shared mobility services analysed

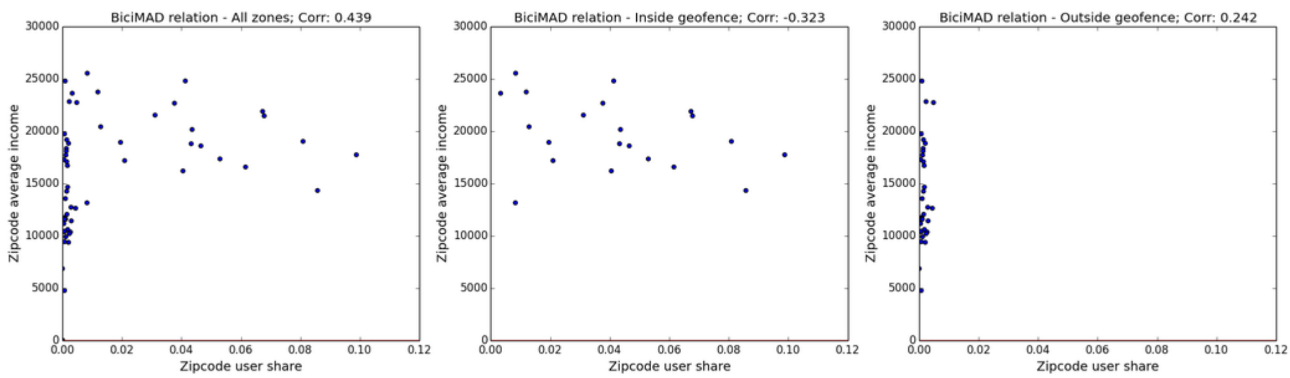


Figure 34. Income against user amount of the population for the BiciMAD service in the cases of all zipcodes (left), zipcodes inside geofence (centre) and zipcodes outside geofence (right)

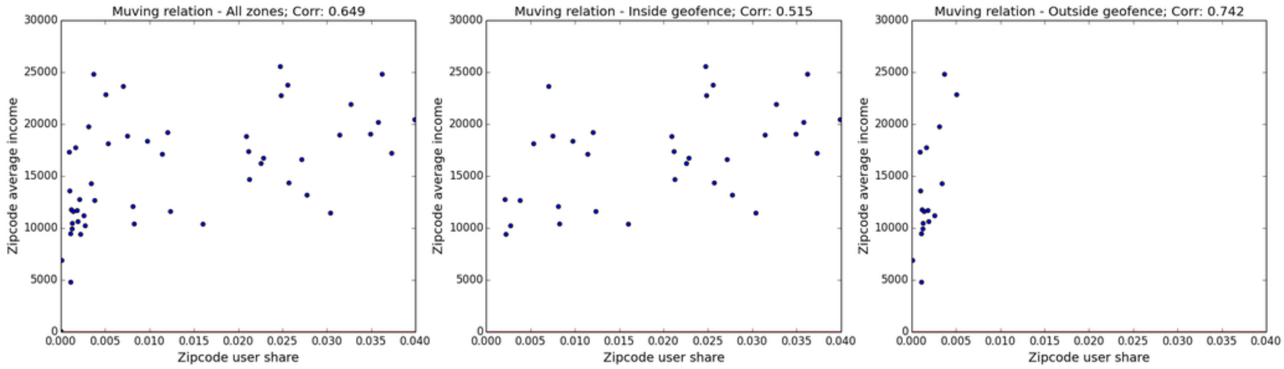


Figure 35. Income against user amount of the population for the Muving service in the cases of all zipcodes (left), zipcodes inside geofence (centre) and zipcodes outside geofence (right)

Interestingly, Figure 35, that shows the same plot for Muving presents larger correlation values between income and user share in the three cases. All this suggests that the Muving service is indeed biased towards higher income users, probably living inside the geofence (city centre). It is particularly interesting to see that the zipcodes with higher income outside the geofence are actually the ones with higher proportion of Muving users among their population (right plot).

In conclusion, these results indicate that while the Muving service fits the expected profile observed in the literature, BiciMAD is a shared mobility service that reaches a wider user-base, both in terms of income and user location (inside/outside geofence), probably due to its cheaper prices and the convenient integration with the regional public transport network through some integrated ticketing measures (discount in the service pass if the user holds a public transport card).

3.1.3.4.3. Control case: experiments using Muving customer database

The previous results have been computed for both services following the trip-based methodology described in Section 2.5.3.2. Nonetheless, Muving has provided MOMENTUM project with a register containing all service users and their home locations, so it is possible to repeat both experiments for Muving user base, both to validate the trip-based methodology along with previous Muving results.

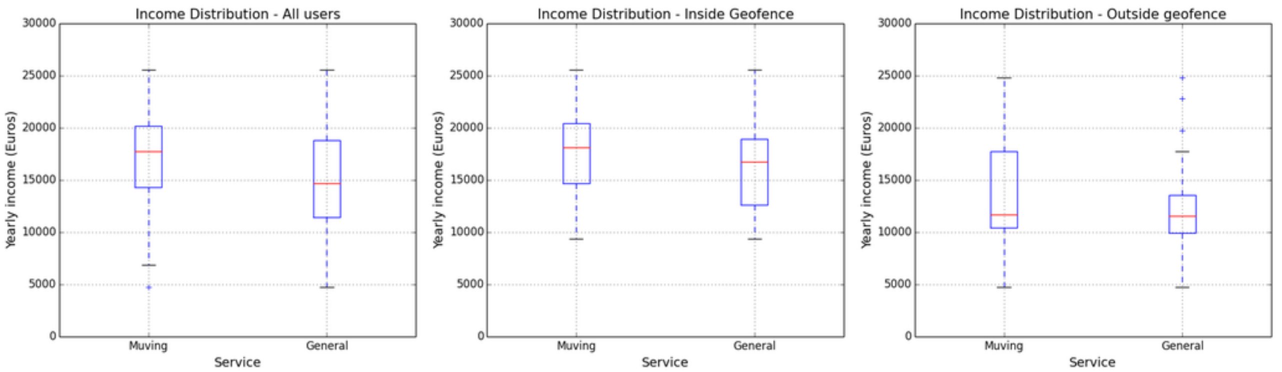


Figure 36. Comparison of Muving user-based income distribution compared with general population income distribution for the cases of all users (left), users living inside the geofence (centre) and users living outside the geofence (right)

In the case of income distributions, the Muving distribution is stable, being the distribution the same from both approaches. These results suggest that the users that performs trips on a daily basis average conform a representative sample of the general user base, at least from the income perspective.

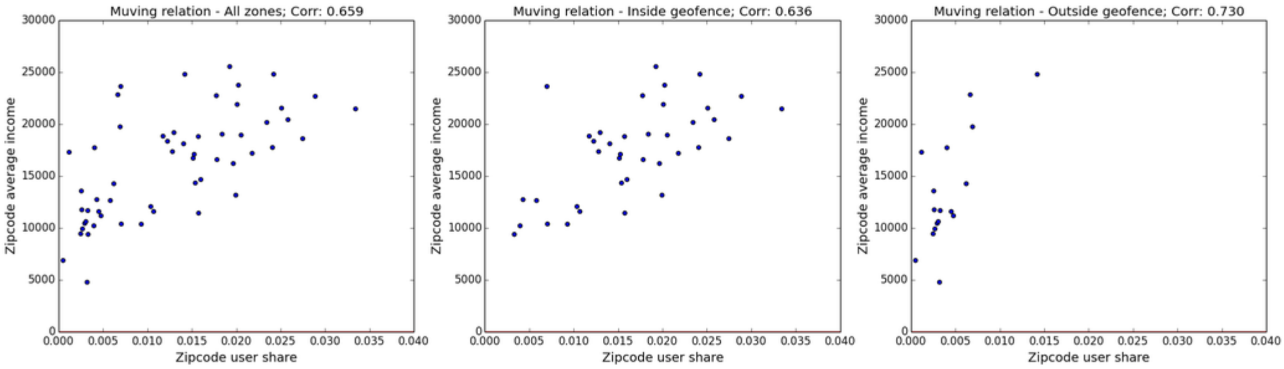


Figure 37. Income against user amount of the population plot for the Muving service (user records) in the cases of all zipcodes (left), zipcodes inside geofence (centre) and zipcodes outside geofence (right)

Regarding the comparison between income and user base, Figure 37 shows some differences with respect to the trip-based case. Still, the trends observed are similar, indicating higher correlation between income and user base, even higher in this case.

3.1.4. Shared-mobility use

3.1.4.1. Shared mobility use patterns from household survey data

The household mobility survey is delivered in four different data aggregations, namely households, individuals, trips and trip legs. For each trip leg, the main mode is identified over an extensive list that includes shared modes. Among all possible modes, we have selected the following ones as shared mobility modes:

- "Bicicleta pública" (Public bicycle)
- "Moto/ciclomotor público" (Public motorbike/scooter)
- "Coche conductor alquiler sin conductor" (Driver of a driverless car rental)

The reader should note that there are some ambiguities in the naming of the modes that could introduce noise into the analysis. As an example, there is a category for "Company bicycle" and another for "Public bicycle"; the latter has been selected since BiciMAD, the most relevant bike sharing mobility service, is explicitly considered as an example of "Public bicycle". Additionally, the category "Driver of a driverless car rental" could include traditional rental companies, even though the number of people renting a car for regular day trips should be insignificant. In order to perform comparison within the survey, any trip that contains at least a leg in any of these three modes is recovered (shared modes users) and compared with all the trips performed in the region (all travellers).

3.1.4.1.1. Number of trips

The Madrid household mobility survey (EDM for short) reports 143 users who declare to use shared mobility modes on their average working day trips, concisely up to 223. Once expanded, there is a total of 10,959 people and up to 16,541.67 trips distributed among shared bike, shared motorcycle and shared car services. The reader should note that the proportion shared mobility services represent conform a very small sample for the total number of survey respondents (85,064) and trips (222,744). Hence, shared mobility user characterization coming from the survey should be carefully considered.

Figure 38 displays the share of each of the three identified modes within all shared mobility users identified.

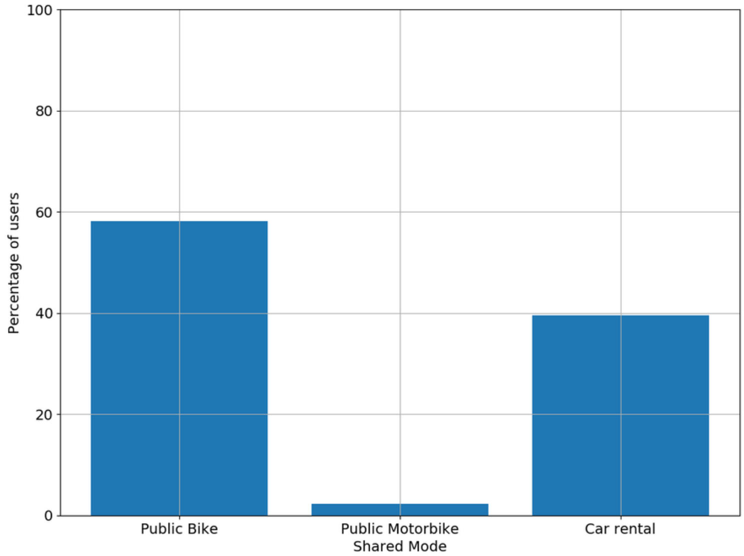


Figure 38. Amount of users of each mode over all the shared mobility users in the region

In the figure, the reader can clearly identify that the majority of shared modes users take bike sharing, followed by car sharing and, much less, motorbike sharing. Specifically, the EDM accounts for 9,611.97 trips in bike sharing, 6,557.28 trips in car sharing and 372.42 trips in motorbike sharing.

This difference might be influenced by the stability of the BiciMAD bike sharing service in the city of Madrid, which was one of the earliest services deployed and is currently one of the most expanded and popular within the city. Furthermore, this service is integrated within the region public transport card system and offers discounts to public transport users to motivate the use of bike sharing in multi-modal trips.

3.1.4.1.2. Trip purposes

One of the most interesting features of the EDM survey with respect to shared modes is the possibility to identify the trip purpose. Figure 39 compares the origin-destination frequency barplots of trip purposes for shared modes and the general population.

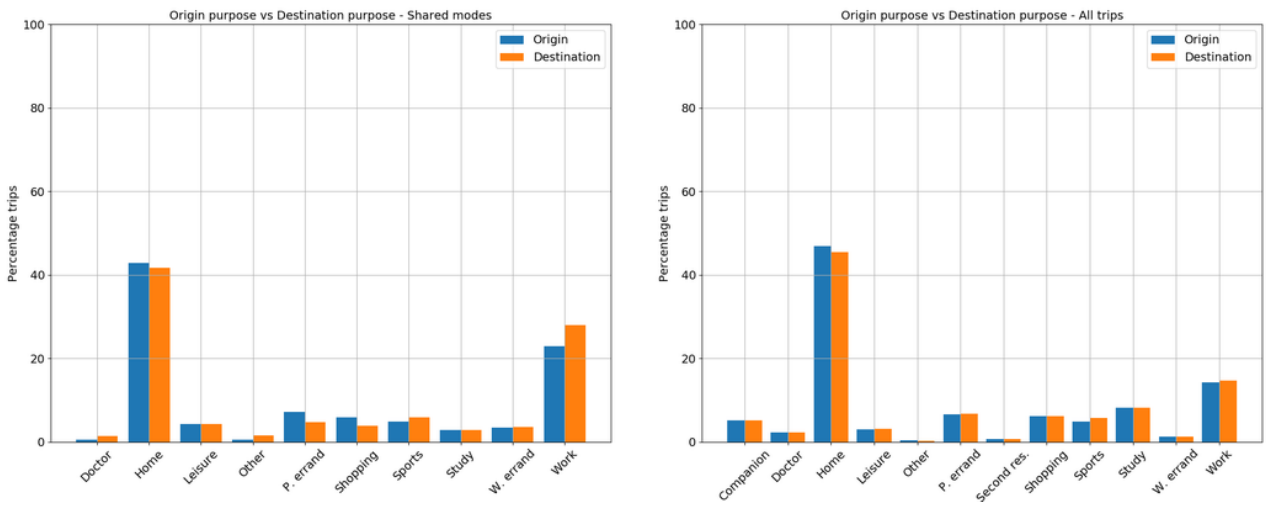


Figure 39. Trip purposes by origin destination of shared modes (left) and general population (right)

The following aspects can be highlighted:

- Work purpose stands out as more frequent among shared mobility trips compared to the total average. The fact that most shared mobility trips are related to the bike sharing service, which is relatively affordable for performing daily home-work trips, it may be causing this result. The outcome follows recent studies identifying home-work trips as the prominent trip type in bike sharing services (Jiao et al., 2020). In contrast, car sharing use has been usually related to other trip purposes (Schmöller et al., 2015). The limited sample size of the trips captured by the household survey makes difficult to extract robust evidences of this from the survey data.
- Home purposes are less frequent for shared mode users.
- The purposes second residence and being someone's companion are not reported for shared mode users, given the limited sample available.
- Leisure and work errand purposes are slightly more frequent as both origin and destination purposes in the case of shared mobility, possibly due to car sharing trips as highlighted in the existing literature (Schmöller et al., 2015).
- The study purpose is significantly less frequent for shared users than the general population, possibly due to the fact that there are some age limitations to the use of the services (14 years old for bike sharing service).

3.1.4.1.3. Private vehicle availability when performing a shared mobility trip

Another question in the survey queries respondents whether they had a private vehicle available for the trips they perform in other modes, which is an indicator for whether the mode has been chosen voluntarily. Among the total 16,541.69 trips where some shared mobility mode is involved the majority of users, up to a 67%, declared having a private vehicle available for those trips.

In contrast, only 42.5% of public transport users acknowledge the availability of a private car for their trips in that mode while for walkers or bikers (private bike), private car availability is slightly larger, 51.1% of all users. Both observations indicate that both modes include a higher proportion of mode users that actually have no alternative means for their trips compared to shared mode vehicles, suggesting that the latter are actively choosing shared modes more often.

Analysing each of the modes separately, 44% of bike sharing users had a vehicle available for the trip which is very similar to the figure of public transport users. Hence, this would be consistent with many of them being regular public transport users that have incorporated shared modes into their trips. On the other side, car sharing users declare having a private vehicle available in all their trips, which has two possible explanations: either the question was misinterpreted by survey respondents or, alternatively, most car sharing users did actively choose that mode even having their own vehicle because they find it more convenient (i.e, avoid paying for parking space or not intending to use the car to come back).

3.1.4.1.4. Multimodal trips including shared mobility legs

The shared mobility adoption analysis conducted in this case study revealed that most users own public transport cards of some kind, even though it is unclear whether they have them as public transport users or to benefit from the associated incentives for some shared modes (BiciMAD). In addition, 22% of the trips that include a shared mode are also part of a multimodal chain.

To further understand whether shared mobility mode users are also public transport mode users, we analyse the modal chains of those trips containing at least a shared mode to determine the preferred combinations. The following Figure 40 shows the most common trip modal chains that involve shared modes and the number of responses obtained for the survey (sample level).

```
[('Rental car no driver', 92),
 ('Public bike', 74),
 ('Public Motorbike', 5),
 ('Metro-Metro-Public bike', 4),
 ('Renfe Cercanias-Public bike', 3),
 ('Public bike-Renfe Cercanias', 3),
 ('Public bike-Bus', 3),
 ('Bus-Public bike', 3),
 ('Public bike-Metro-Metro', 3),
 ('Bus-Metro-Public bike', 2),
 ('EMT-Metro-Public bike', 2),
 ('Public bike-Renfe Cercanias-Metro-Metro', 2),
 ('EMT-Public bike', 2),
 ('Urban bus-Renfe Cercanias-Metro-Public bike', 1),
 ('Public bike-Metro-Renfe Cercanias-Urban bus', 1),
 ('Public bike-Metro-Bus', 1),
 ('Urban bus-Public bike', 1),
 ('Public bike-Urban bus', 1),
 ('Public bike-Metro-Renfe Cercanias-Private car driver', 1),
 ('Public bike-Renfe Cercanias-Metro', 1)]
```

Figure 40. Most frequent modal chains and their frequency of appearance in the surveys

The figure shows that multimodal trips mostly involves bike sharing and public transport, which is the actual behaviour in all cases. This suggests that bike sharing services have been successfully integrated inside the modal options available in the region enabling the public transport network to reach further through them. In contrast, car sharing and moto sharing services would be acting as more door-to-door services not integrated within the public transport network, possibly due to their convenience and flexibility.

3.1.4.1.5. Trip distribution

The survey provides the origin and destination areas of the trips given by each respondent, so it is possible to track most popular origins and destinations of shared mobility modes. Figure 41 depicts the coloured map of most frequent origins whereas Figure 42 depicts the same map but for most frequent destinations.

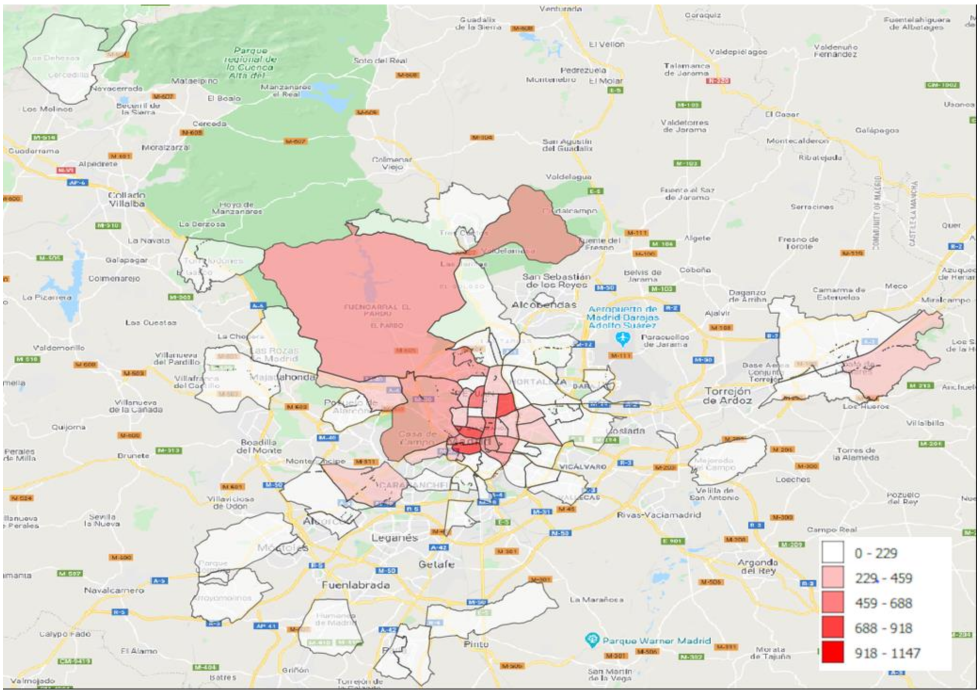


Figure 41. Colored map of most frequent origin zones for shared modes

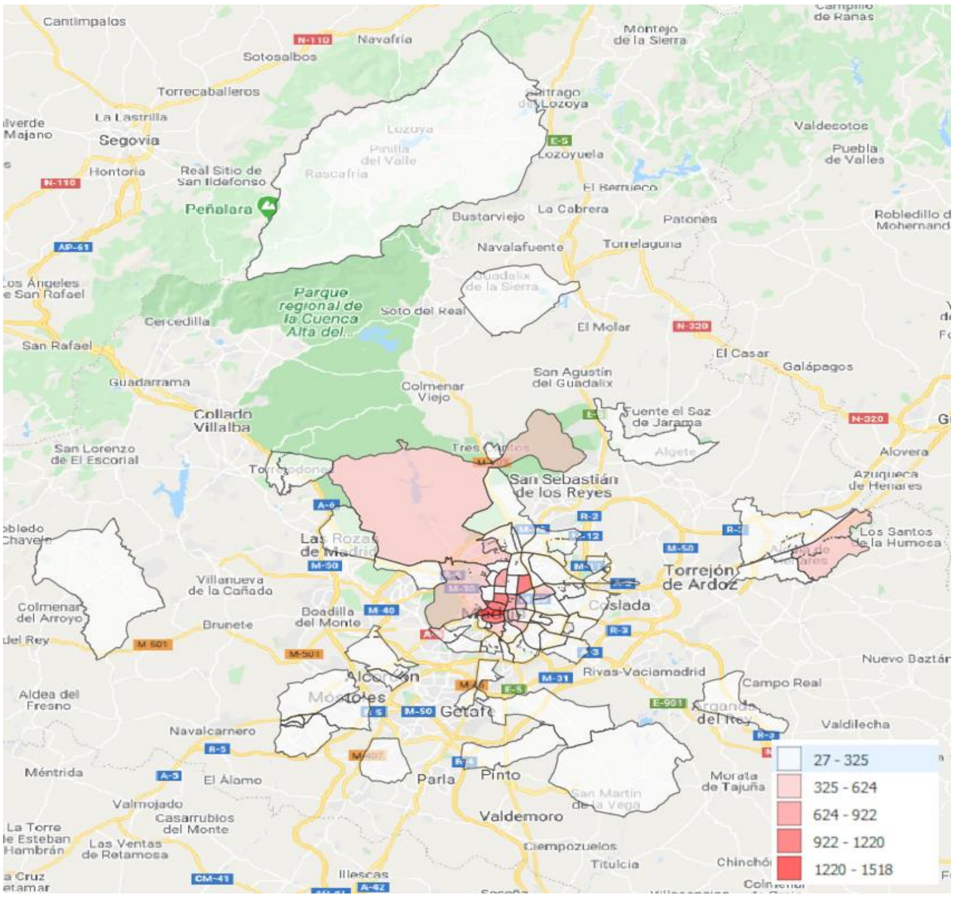


Figure 42. Colored map of most frequent destination zones for shared modes

Both images show that most trips prevail within Madrid's centre region ("Almendra central"), which makes sense since it is the basic area of coverage for the majority of shared mobility services operating in the city. Despite, there are other interesting origin and destination areas, mainly in the east and south parts of the city's metropolitan area. The reader should bear in mind that the trips in this analysis include shared mobility modes that are used either for the entire trip or just as a leg in a multi-modal trip, which explains the reason for the appearance of zones outside of the regular area of operation.

Actually, the amount of trips to or from origin and destination areas outside the "Almendra central" area belong to different cities in the metropolitan area connecting with Madrid centre through the Cercanías train network or inter-city buses. Such services have a coarser granularity within Madrid and therefore, their combination with bike-sharing services facilitates the accessibility to wider areas within the main areas of the city.

Most of the users performing trips originated or destinated inside the metropolitan area tend to show roundtrips, that is, they show the same modal chain to go and come back within the same day. Concisely, 44.9% of the survey respondents that declared using shared mobility perform more than one of their trips using shared mobility and, of those, 73.4% do roundtrips (that is going from A to B and coming back from B to A after in the same day).

3.1.4.1.6. Conclusions

The limited availability of sample trips in the household mobility survey implies that only certain aggregated use analyses are meaningful, but some conclusions can be found at this level. First, the relevance of bike sharing among total shared mobility trips implies that home-work purpose accounts for a large share of the trips made with these modes. Second, shared mobility users have declared having available a private car for their trips 67% of the times, much above public transport users (42.5%) and walkers (51.1%). This suggests that shared mobility services may be found more attractive than public transport services for private car owners. Finally, it has been observed that some shared modes are used as part of larger modal chains involving public transport. Specifically, 22% of bike-sharing trips are performed along with public transport modes unveiling a certain degree of complementarity between both. No multimodal chains were found for car sharing users. Additionally, it has been observed that long trips to/from the metropolitan area to Madrid's city centre that include shared modes are frequently roundtrips within the same day.

3.1.4.2. Shared mobility use patterns from operation data

3.1.4.2.1. BiciMAD shared mobility service

BiciMAD data is available at the service web portal for public download. The data available includes:

- Usage data by months, which contains all trip records registered with starting time, duration and starting and finishing stations. Each trip contains the ID of the user that takes the bike as well as their age, the zipcode of their home location and the type of user they are (frequent, occasional or service worker).
- State of all the stations in the city, including position and number of available bikes periodically.

A major downside with BiciMAD data is that, with the aim of preserving privacy, user IDs are anonymized and changed on a daily basis, so it is not possible to perform user longitudinal analysis for time periods larger than a day. As a result, only daily usage can be analysed and for larger time periods "user-days", including all the trips of a user in one day, will be considered rather than complete users with their longitudinal trip analysis.

The dataset studied for the BiciMAD service comprises all trips registered during the following months: October 2018, December 2018, February 2019, April 2019, June 2019, July 2019, October 2019, December 2019 and February 2020.

3.1.4.2.1.1. Daily trips per user-day

Figure 43 depicts the usage histograms of the service by years (including only the study months) as the percentage of user-days that take the service a number of times.

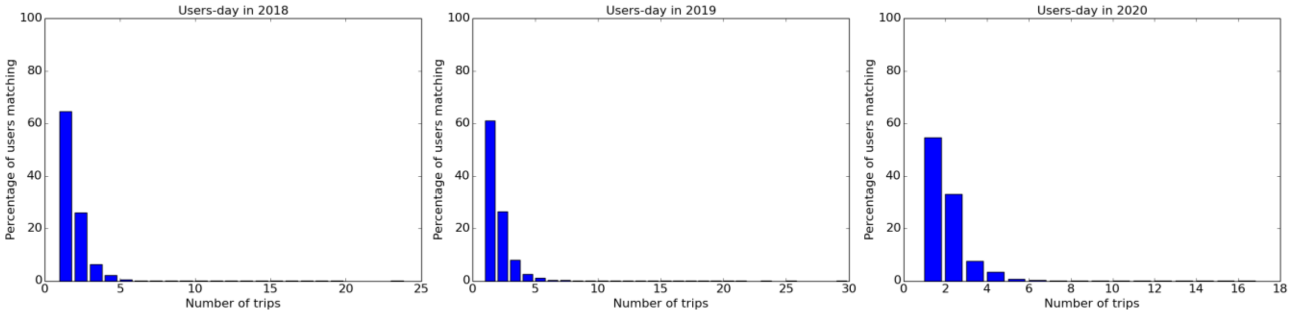


Figure 43. BiciMAD usage histograms for 2018 (left), 2019 (centre) and 2020 (right)

The figure shows a similar pattern for the three years considered, where most of the time the majority of users take the service for a single trip in a day, suggesting that a good part of the usage of this service is for non-recurrent and possibly non-frequent trips.

Nevertheless, the amount of user-days that perform two or three trips a day is still considerable, comprising nearly 40% of all users. This percentage has slightly grown over the years. Regardless, the use of the service for more than 5 trips is very uncommon, as can be observed in the three graphs where the histogram drops to almost 0 when that number of trips is reached.

3.1.4.2.1.2. Trip distance distribution

Figure 44 depicts the regular and the cumulative distribution of trip distances in kilometres for the BiciMAD service.

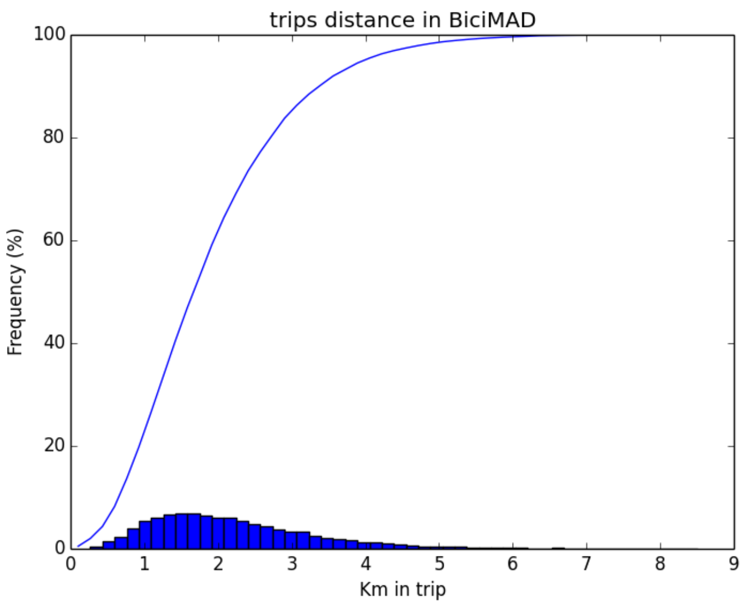


Figure 44. Cumulative distribution and histogram of trip distances for the BiciMAD service

Figure 44 shows a fast increase of the trip counts, especially in the range between 1 and 3 km, where the increasing rate starts decreasing. In general, it might be observed that BiciMAD trips are short, with most frequent distances within the aforementioned range even though there are trip distances of up to 8 km. The average and median distance values are 2.03 and 1.83 km respectively.

3.1.4.2.1.3. 3.1.4.2.1.3 Trip duration distribution

Regarding trip duration, Figure 45 displays a histogram for the number of minutes taken per trip and their frequency for all the trips observed in BiciMAD dataset.

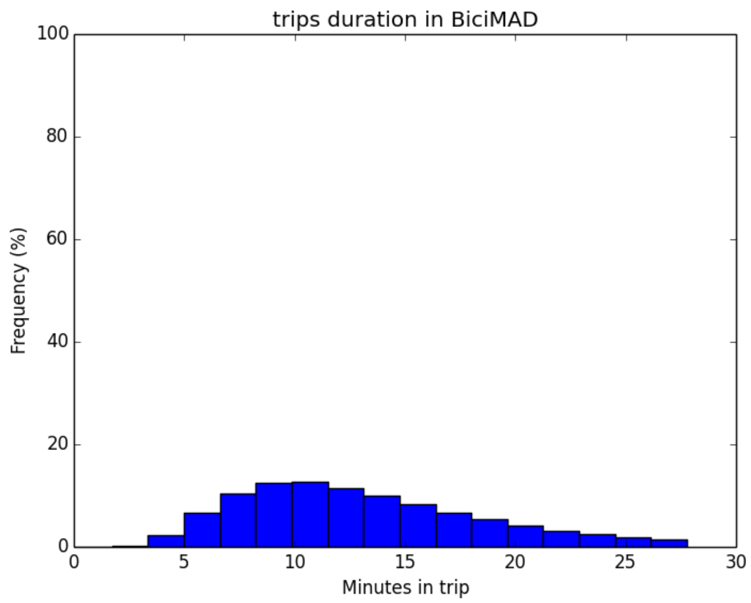


Figure 45. Distribution of trip durations in minutes

As observed in the figure, the most frequent trip durations occur in the range 7-13 minutes, which is consistent with trip distances of 2-3 km. The average and median trip duration are 19 and 11 minutes, respectively. It is also worth to mention that trips shorter than 5 minutes are very infrequent, similar to trips longer than 30 minutes, possibly due to the fare scheme, which is fixed for the first 30 minutes.

3.1.4.2.2. Muving shared mobility service

Muving trip information is available for this case study thanks to the company collaboration agreement for the MOMENTUM project. The dataset under study comprises all trips registered by the service during the following months: October 2018, December 2018, February 2019, April 2019, June 2019, July 2019, October 2019, December 2019 and February 2020.

The dataset contains a register of all the trips observed in the service with their origin, destination, beginning and ending times as well as the ID of the user performing the trip along with some of their sociodemographic characteristics, namely age and home location zipcode. In this case, user IDs are the same through time, so it is possible to track users through time.

3.1.4.2.2.1. Daily trips per user-day

Figure 46 depicts the usage histograms of the service by years (including only the study months) as the percentage of user-days that take the service a number of times.

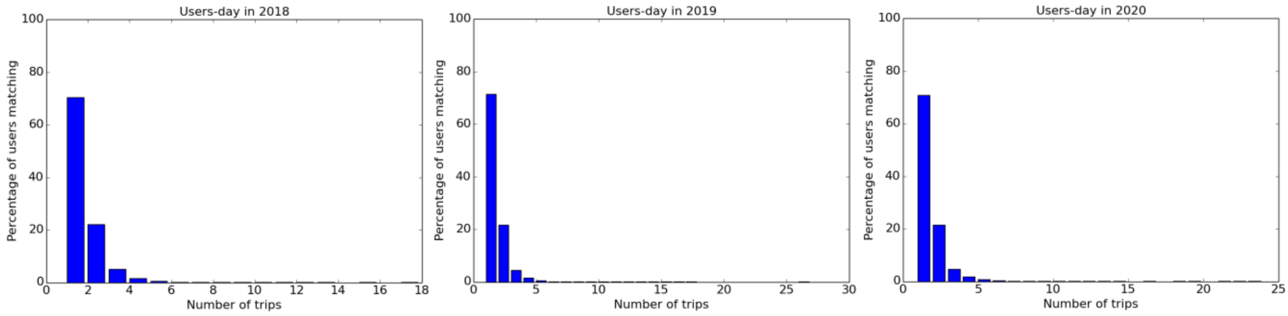


Figure 46. Muving usage histograms for 2018 (left), 2019 (centre) and 2020 (right)

As in BiciMAD, the pattern is the same in the three years, where most users take the service for a single trip in the day. In this case, it can be observed that single daily uses in the service are proportionally larger than BiciMAD's, suggesting that less users from this service take it for regular trips. Now, the amount of user-days that perform two or three trips a day is below 30%. The values have remained stable across the years.

3.1.4.2.2.2. Monthly trips per user

Since users can be analysed through time, Figure 47 provides the service usage histograms on a monthly basis: for each month, the number of days a user travels at least once using Muving is recorded and aggregated to obtain the percentage of users taking the service for different number of days.

In general, the figure highlights that the majority of users, nearly 60%, only take the service once a month. Note that users that have trips in other months are not included, so the total of users displayed at each month comprise exclusively all those user IDs that report at least one trip within the month.

Interestingly, the proportion of one day users decreases during the months of June and July 2019, which could be an indicator of more frequent use during summer months, when weather is typically better. On the contrary, frequent use of the service is lowest in December 2019, when the total amount of users that take the service only once surpasses 60%.

In addition, the number of users that take the service on a more regular basis, for instance once a week or more, is lower than 10%. Hence, it appears that most users perform a trip without regularity, possibly selecting the service as their mode of transport for very specific non-frequent events.

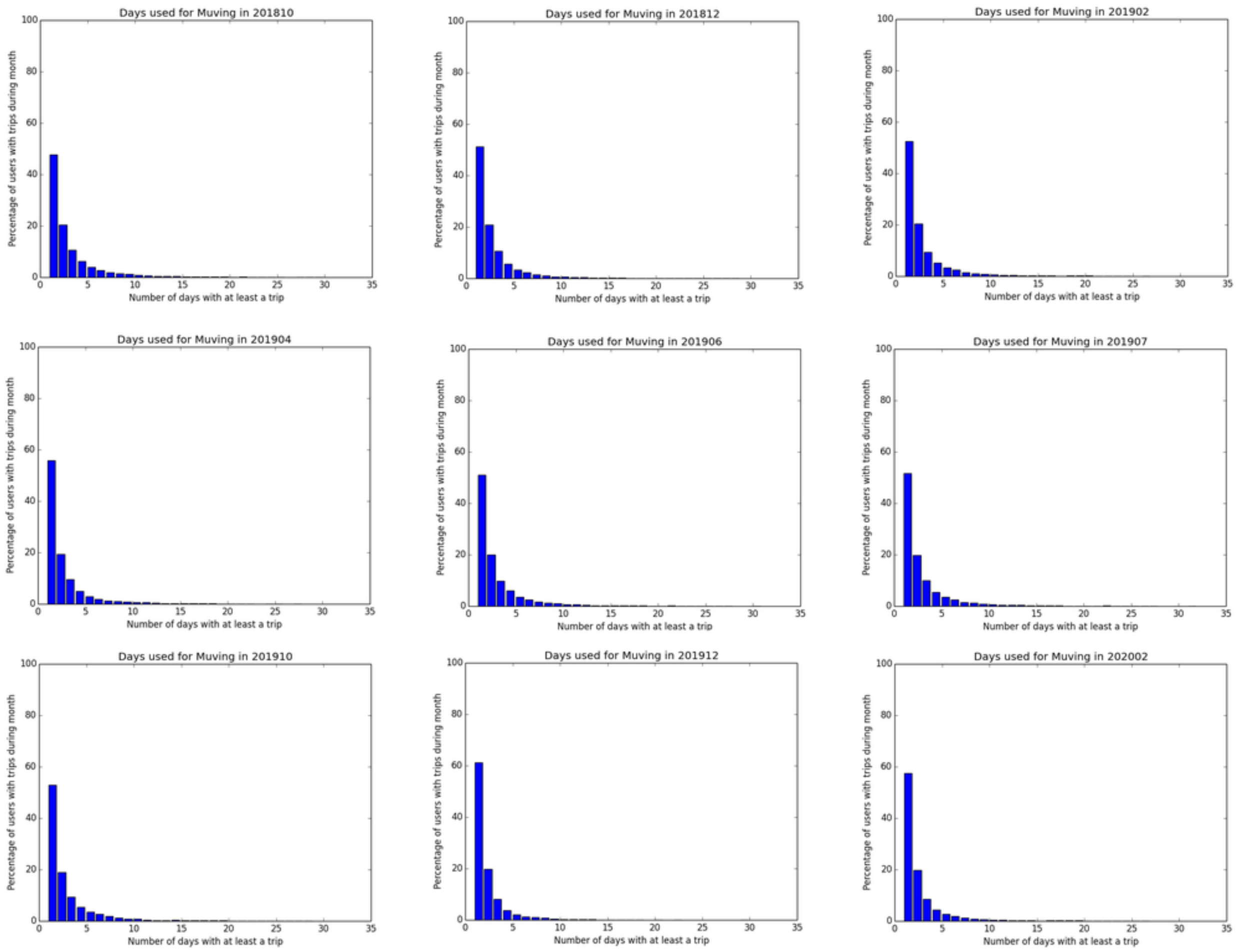


Figure 47. Monthly histograms of the days using the service and the amount of users

3.1.4.2.2.3. 3.1.4.2.2.3 Trip distance distribution

Figure 48 depicts the cumulative distribution of trip distances in kilometres for the Muving service. This figure shows a much faster increase in terms of distance, being the largest increase in distances between 1 and 4 km. The trip counts reach almost 100% when the distance is approximately 8 kilometres. In general, distances are slightly larger than in the case of BiciMAD; nonetheless, the distance range is similar and the average and median distances are 2.5 and 2.15 km, which is moderately larger than in the BiciMAD case.

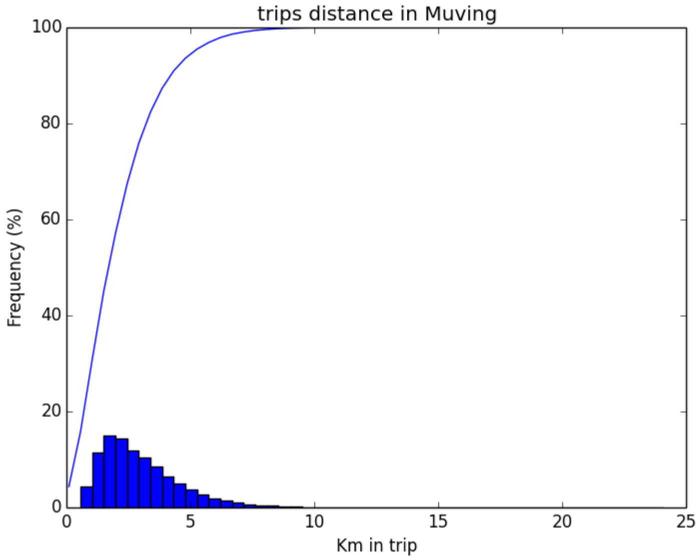


Figure 48. Cumulative distribution and histogram of trip distances for the Muving service

3.1.4.2.2.4. Trip duration distribution

Figure 49 displays the histogram of the trip duration distribution in minutes.

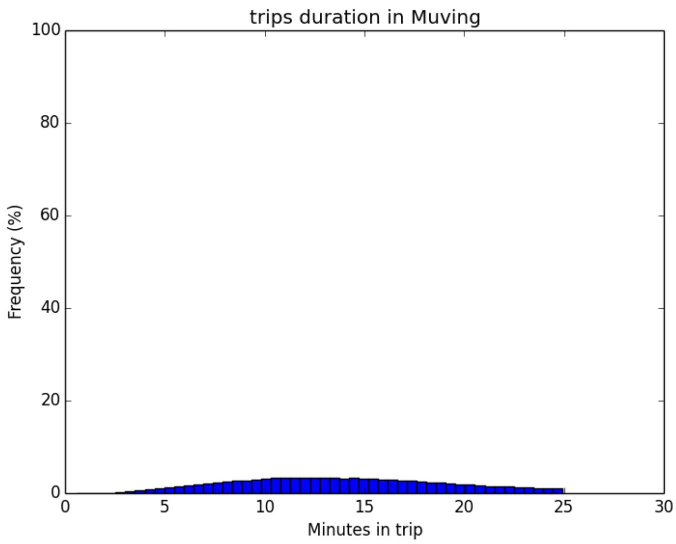


Figure 49. Trip duration distribution histogram

The figure shows a very flat distribution, which indicates trips durations are nearly uniformly distributed, mainly between 5 and 22 minutes with outliers. The average trip duration is 16 minutes whereas the median duration 14 minutes.

Interestingly, the trips performed in the Muving service are faster than those performed in BiciMAD, in spite of travelling a slightly longer distance. In any case, the difference in terms of distance is small compared to the difference in terms of features of BiciMAD's e-bikes with respect to Muving motorbikes.

3.1.4.2.3. Trip purpose characterisation through fusion with mobile network data

Any OD matrix provides an extensive and representative characterization of mobility patterns in different modes, such as a shared mobility service or even the entire travelling population within an area. Once shared mobility services are computed, they can be compared with the OD matrix for general mobility, which has been obtained through mobile network data.

Mobile network-based OD matrices provide an estimation of trips for a given day and might be separated according to different sociodemographic criteria, like gender or age, as well as other criteria, such as trip purpose (either at origin or destination). Generally, all these segmentations are seldom available within shared mobility operation data.

In this light, the comparison of trends between shared mobility OD matrices and particular segmentations of the general mobility OD matrix could provide further insights on relevant aspects, such as the trip purpose. This is particularly interesting to overcome one of the drawbacks of shared mobility operation data, which is the lack of trip purpose information. In addition, demographic characteristics can be used to help to filter or to separate the population according to the most relevant demographic characteristics. As a result, the comparisons between a shared mobility service and general mobility through this section are based on the following segmentations:

- **E1_H**: Age group 1 (between 20 and 44 years old) and destination purpose home.
- **E1_W**: Age group 1 and destination purpose work.
- **E1_O**: Age group 1 and destination purpose other frequent activities.
- **E1_NF**: Age group 1 and destination purpose non-frequent activities.
- **E2_H**: Age group 2 (between 45 and 64 years old) and destination purpose home.
- **E2_W**: Age group 2 and destination purpose work.
- **E2_O**: Age group 2 and destination purpose other frequent activities.
- **E2_NF**: Age group 2 and destination purpose non-frequent activities.

The selection of age groups covers all the people that are most likely to be a user of a shared mobility service, since younger people are not typically allowed to use them on their own (due to being underage or not having the proper licenses) whereas the literature shows that the use of shared mobility services by the elderly is negligible.

3.1.4.2.3.1. BiciMAD service

Figure 50 shows the results of the regression analysis of the bike sharing OD matrix compared to the different OD matrices from overall mobility segments according to age and trip purpose.

The patterns observed in Figure 50 are very clear in the distribution of trips, specially attending to destination purpose: the destination purpose for which BiciMAD indicates a larger correlation is the other-frequent purpose, followed by home purpose, non-frequent, especially around weekends and, finally work purpose. Actually, correlations on weekends for the non-frequent purpose are higher than those with the Home purpose. It has also to be noted that the correlation with work and home purposes has increased in the last analysed months. The two groups of age ranges selected show small differences among them.

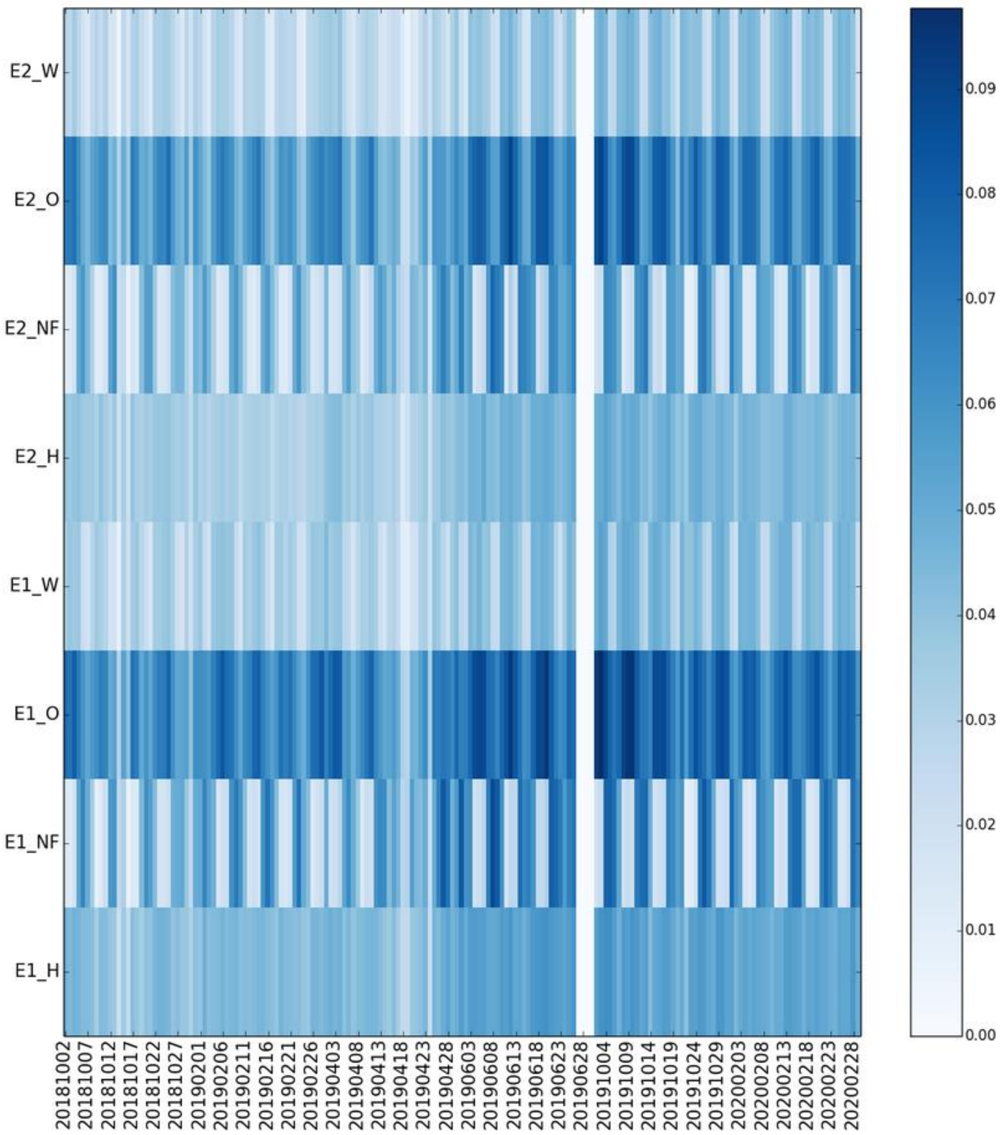


Figure 50. Correlation between BiciMAD OD matrix and a selection of general OD matrix segmentations for all the days in October 2018 and February, April, June and October 2019 and February 2020

3.1.4.2.3.2. Muving service

Figure 51 shows the results of the regression analysis of the moto sharing OD matrix compared to the different OD matrices from overall mobility segments according to age and trip purpose.

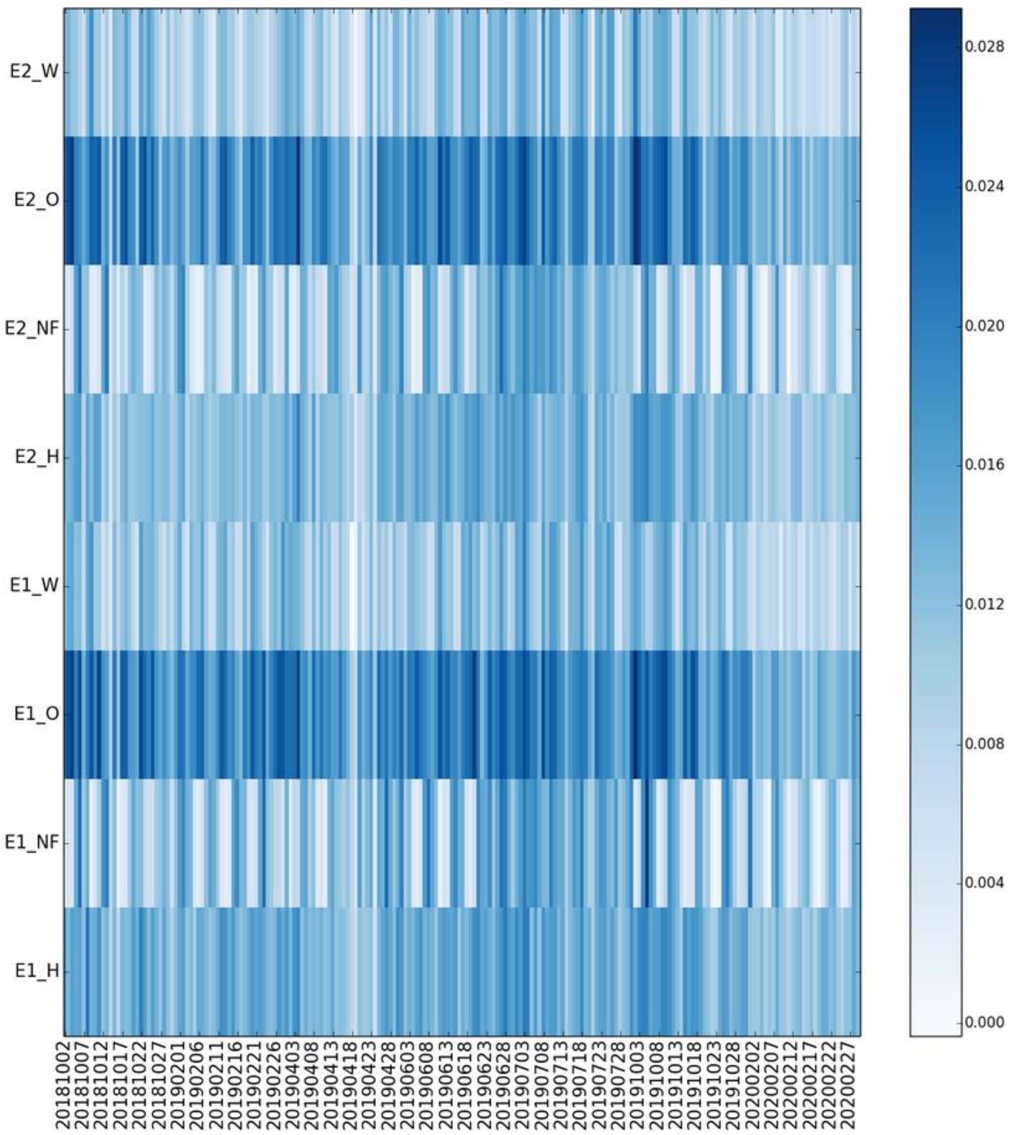


Figure 51. Correlation plot between Muving OD matrix and some segmentations of the general mobile network OD matrix for the months of October 2018 and February, April, June and October 2019 and February 2020

The patterns observed in the figure are similar to those of BiciMAD, even though there is more noise through them. The other frequent destination purpose is the most relevant, followed by home and non-frequent, which again has a marked pattern during weekends complementary to the other frequent pattern that gets smaller during weekends. Absolute correlation values are also lower than in the case of BiciMAD. One expected outcome would be a sharper difference between both systems in terms of correlation with home-work flows. As mentioned above, the literature suggests that bike sharing services are more used for home-work trips that other shared modes. While it is true that the correlation in the case of BiciMAD is higher than in the case of Muving (average of 0.04 vs 0.012), this increase is not higher than the one observed in other purposes. Therefore, this approach deserves further research.

3.1.4.2.4. Conclusions

The data fusion exercise conducted with shared mobility operation data and mobile network data provides some interesting insights. The data analysed from both shared mobility services has shown that the most prevalent use of the service is for non-recurrent trips and with no regularity, being the frequency of usage typically low. In fact, the most correlated trip purposes from the general OD matrix are the other-frequent for weekdays and non-frequent activities for the weekends, which are consistent with non-regular usage.

Nevertheless, it is worth noting that BiciMAD service shows more regular use with respect to Muving, especially observed in the lower amount of user-days that only perform a single trip. In fact, it has been observed that more than half Muving users take the service only once a month. Correlation with the general OD matrix segmentations is also higher in the case of BiciMAD, although it is not found to be specifically higher for the work purpose

3.1.4.3. Shared mobility role in restrictions to private vehicles

One of the claimed benefits of emerging mobility solutions is that they provide an attractive alternative to the use of private vehicles. The assessment of this claim is crucial for evaluating the contribution of shared mobility and other mobility innovations to urban sustainability. As a consequence, the identification of modal shifts from private vehicles to shared mobility services has been claimed to be an urgent research question (Shen et al., 2018). Several attempts are taking place to explore this issue. In the context of bike sharing, the recent reviews show a limited impact of these schemes in private vehicles use levels (Ma et al., 2020). For instance, Fishman et al. (2014) quantifies this for several cities in the world after a large survey to users, limiting the number of users who would have performed their trips by car from 2% in the case of London to 21% in the case of Brisbane. In the case of car sharing, recent results based on stated preference surveys also suggest that the potential for modal shifts is fairly limited (Liao et al., 2020). Following this, it has been modelled how the extension of MaaS frameworks may increase the shift from car use, thanks to a more direct individual perception of the private vehicle costs in comparison to other modes (Becker et al., 2020).

While most studies rely on surveys to users, the exploitation of longitudinal data sources to analyse this substitution effect has not been much explored. The use of the data collected by operators is common among shared mobility demand characterisation, but the data do not reveal the mode choice patterns of the users beyond the service or before its implementation. Most household surveys are cross-sectional and not longitudinal, given their high costs. Nevertheless, the still limited modal share of these services implies that surveys are able to capture only a small sample of users. Passively collected data sources such as mobile network data have emerged as an alternative to perform such longitudinal studies and describe the evolution of overall mobility patterns. However, one of the main drawbacks of mobile phone data is that its spatiotemporal resolution normally limits the ability to identify mode choice patterns in urban and metropolitan areas. In this context, the limitations of each data source can be partially alleviated by combining them in the analyses. The increasing availability of several data sources characterising mobility demand patterns is an opportunity to explore this option.

Madrid case study in the MOMENTUM project offers a particularly appropriate framework for this analysis, due to (i) the large number of shared mobility services established in the city, and (ii) the recent implementation of an Urban Vehicle Access Regulation (UVAR) zone in the city centre. First, the wide range of shared mobility services available enriches the analysis by providing insights about how different shared modes can be an alternative to private vehicles. Second, the availability of several longitudinal data sources (mobile phone data, traffic count data, shared mobility operation data, etc.) before and after the implementation of the UVAR zone in November 2018 enables an exploration of the dynamic choice patterns in a natural experiment. Following this, the first research question included in the Madrid case study of the MOMENTUM project is to what extent is shared mobility used as a substitute to private vehicle trips, in relation to the implementation of UVAR zones. A series of experiments have been designed to shed light on this aspect, combining the available data sources. First, traffic counts data are used to assess the impact of the UVAR zone implementation in private vehicle demand. Second, the use of mobile network data allows to identify if the potential decrease observed in private vehicle demand is caused by modal shifts from private vehicle to other modes or by activity relocation effects due to the worse accessibility by car (dissuaded demand). Third, shared mobility operation data reveals if these modes were capable of absorbing part of the private vehicle demand in case that modal shifts have happened due to the implementation of the UVAR zone.

3.1.4.3.1. Background

As it has been the case for other European cities, Madrid implemented an UVAR zone called 'Madrid Central' in its city centre in November 2018. This measure was part of the 'A-Plan' tackling air pollution in the city, whose NOx emission levels are higher than international standard limits. The area has an extension of 472 Ha (Figure 52), and restricts access/egress traffic and through traffic. Private cars are only allowed if they prove to have low emission levels, if they are owned by a resident in the UVAR zone or if they use an off-street parking facility. Similar rules apply to motorbikes, except that they have unlimited access from 22pm until 7am. The implementation of the restrictions has been accompanied by several traffic calming and capacity reduction measures in the main streets across the area, such as Gran Via or Atocha. The following events related to the UVAR zone and traffic calming measures are relevant for the longitudinal analysis:

- The capacity reduction in Gran Via street from December 2017.
- The first implementation of the UVAR zone on 30th November 2018.
- The beginning of the UVAR zone fines and penalties on 16th March 2019.
- A temporal suspension of the restrictions from 1st July 2019 to 8th July 2019.

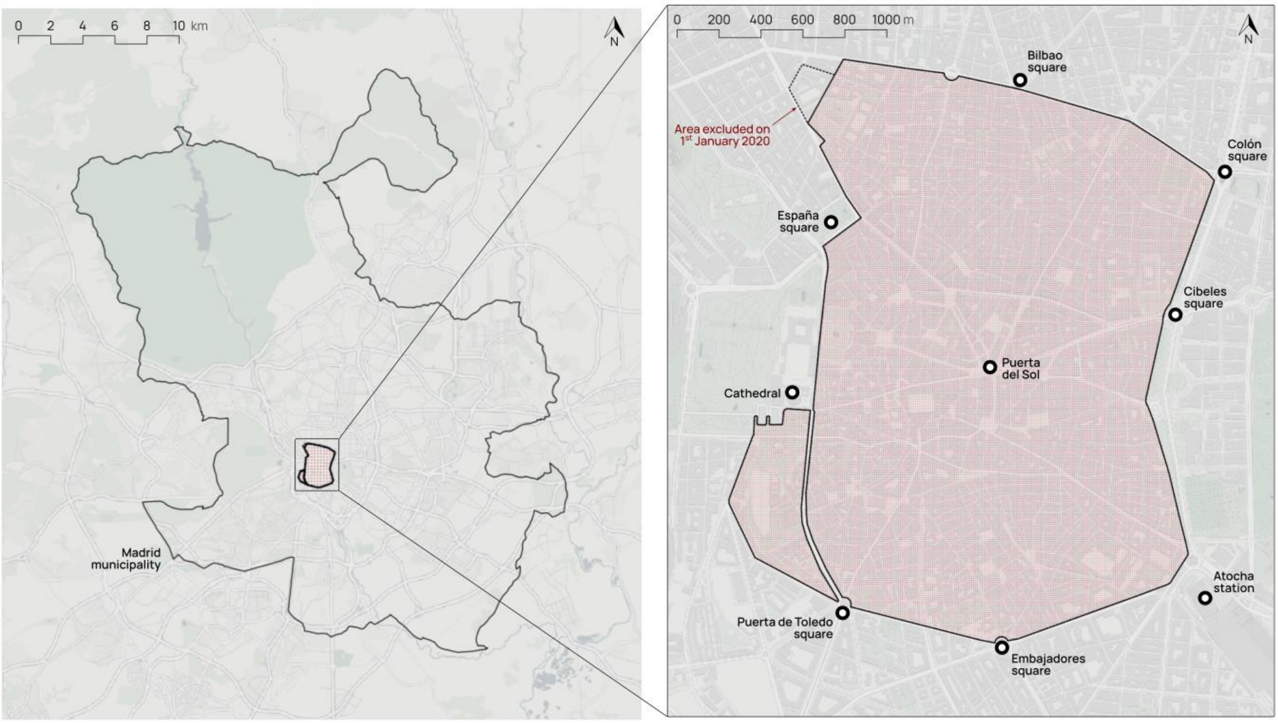


Figure 52. UVAR zone 'Madrid Central' location and perimeter

The implementation of the UVAR zone has happened while many shared mobility operators were deploying services in the city. The current available services (September 2020) include 2 e-hailing providers, several round-trip car sharing providers, 5 one-way car sharing providers, a station-based public bike sharing system, 3 free-floating bike sharing providers, 4 moto sharing providers and 14 e-scooter sharing providers. A complete review of the shared mobility services in Madrid can be found in Arias-Molinares and García-Palomares (2020). As it has been discussed in the rest of this section, MOMENTUM case study is based on the data from the station-based public bike sharing system and from a moto sharing service (Figure 53).

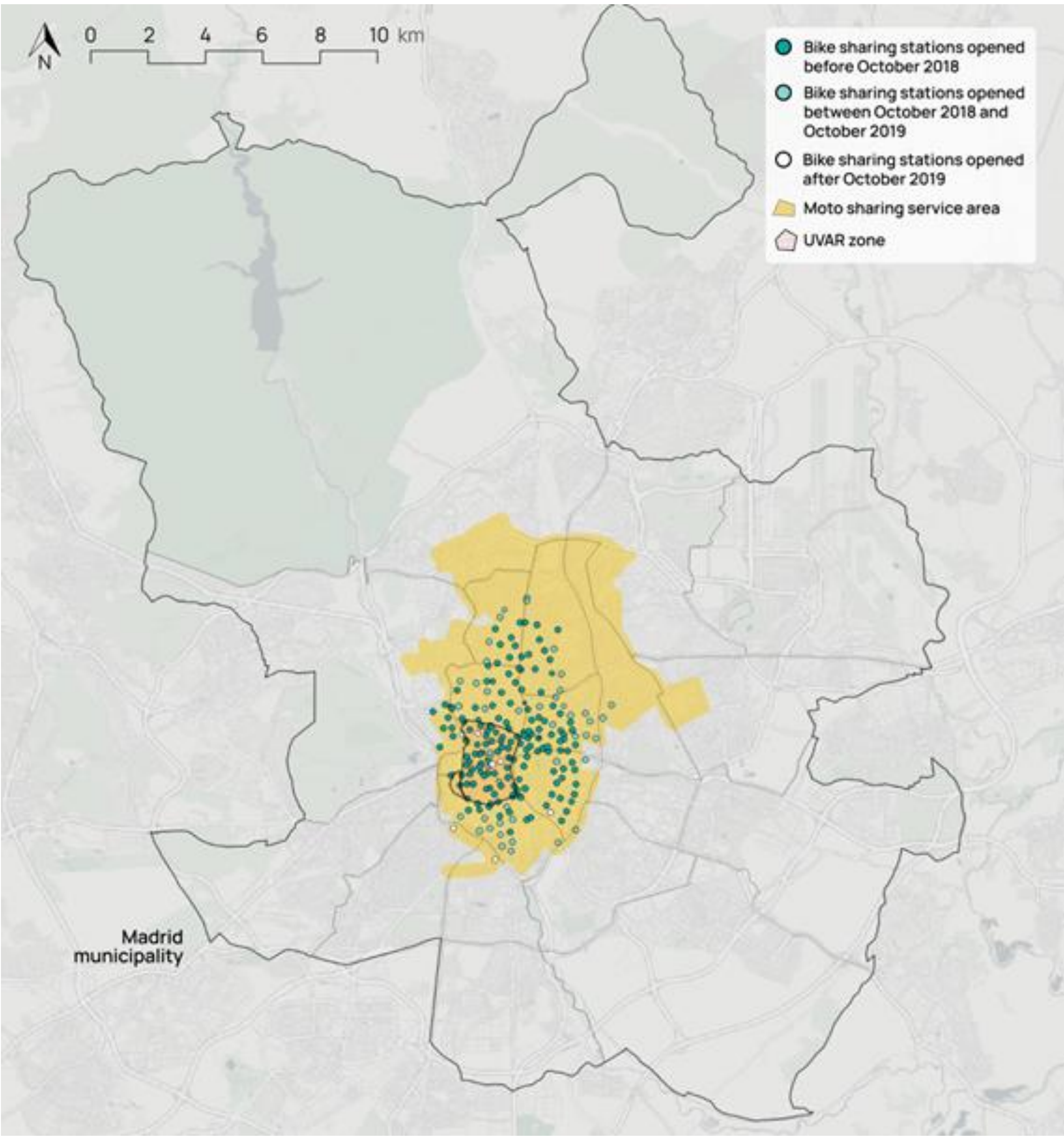


Figure 53. Spatial distribution of bike sharing service and moto sharing service supply

3.1.4.3.2. Impact of the UVAR zone in private vehicle demand

The objective of the first experiment is to assess the impact of the UVAR zone implementation in terms of private vehicle demand. The measure restricted both through traffic (origin and destination outside the UVAR zone) and connection traffic (origin or destination within the UVAR zone). The outcome of the experiment must show if this measure caused a reduction in connection traffic.

3.1.4.3.2.1. Aggregated traffic indicators

First, the aggregated indicators provided by Madrid City Council about traffic evolution in different parts of the city are reviewed. According to these indicators, the total traffic in the city slightly dropped by 2% in 2018 and 3.5% in 2019 compared to 2017 levels (Table x). In parallel, the traffic in the city centre experienced a much larger reduction, especially between 2017 and 2018. The reduction achieved in 2018 was about 11.5% compared to 2017 levels. The additional impact of the UVAR zone in this indicator was rather limited, since the traffic in 2019 would have only decreased 2% from 2018 levels according to this indicator.

Area	2018 vs 2017	2019 vs 2017	2019 vs 2018
City centre	-11,5%	-13,1%	-1,8%
City total	-1,9%	-3,4%	-1,5%

Table 8. Yearly traffic variations in the city centre compared to total in the city

Given that traffic counts are prone to measurement errors, these types of aggregated indicators are used to prevent the influence of outliers in the analysis. However, they are very sensible to sensor locations. For instance, has to be noted that the "city centre" indicator includes traffic counts outside the UVAR zone (e.g. Princesa street). Traffic diversions can also have a strong impact in the indicator if several of the aggregated counts are affected by them, as the Figure 54 shows for the last months of 2019. Hence, a closer analysis is needed for quantifying the influence of the UVAR zone in the traffic volumes. The availability of fine-grained data from traffic counts enables this analysis, as explained in Section 2.

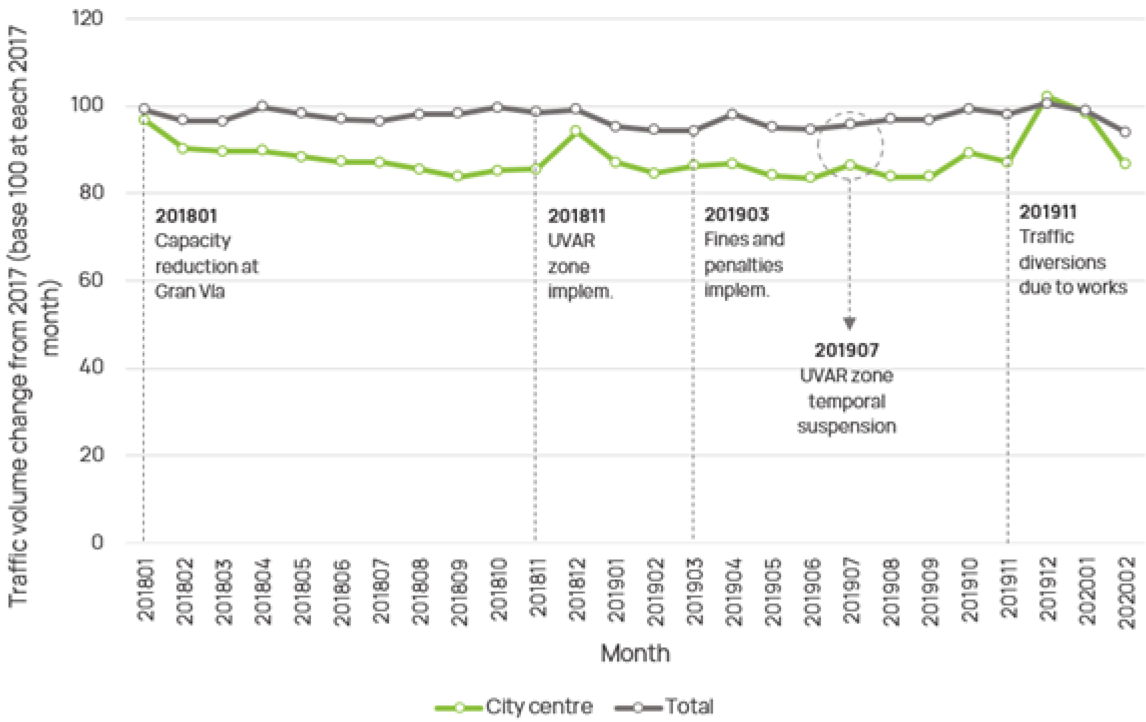


Figure 54. Traffic evolution from 2017, according to the aggregated indicators that Madrid City Council provides from permanent traffic counts

3.1.4.3.2.2. Analysis of detailed traffic counts data

The raw traffic counts data from those points within the UVAR zone and its perimeter are processed to check the evolution of traffic flows before and after the implementation of the measure. Madrid City Council has 101 traffic measurement points in the entrances, exits or within the UVAR zone and 90 traffic measurement points in the streets that shape the perimeter of the UVAR zone. These points include permanent traffic counts (magnetic loops) and eventual traffic count locations, where the traffic is measured with a given periodicity or a pneumatic tube has been installed to make a specific measurement before an urban intervention.

Some traffic counts are affected by street works, as it is the case of the Gran Via street or part of the streets in the perimeter of the UVAR zone, during the construction of cycling lanes. Nevertheless, it is possible to extract a sample of 28 traffic counts along the perimeter of the UVAR zone with enough data continuity from 2017 to 2019. In the case of the entrances and exits to the UVAR zone, it is possible to extract a sample of 13 traffic counts with enough data continuity for the analysed period. These counts are evenly distributed across the entrances and exits to the UVAR zone, as well as along and its boundaries. Figure 55 shows the location of the traffic counts, and Figure 56 shows the data availability in the selected counts.

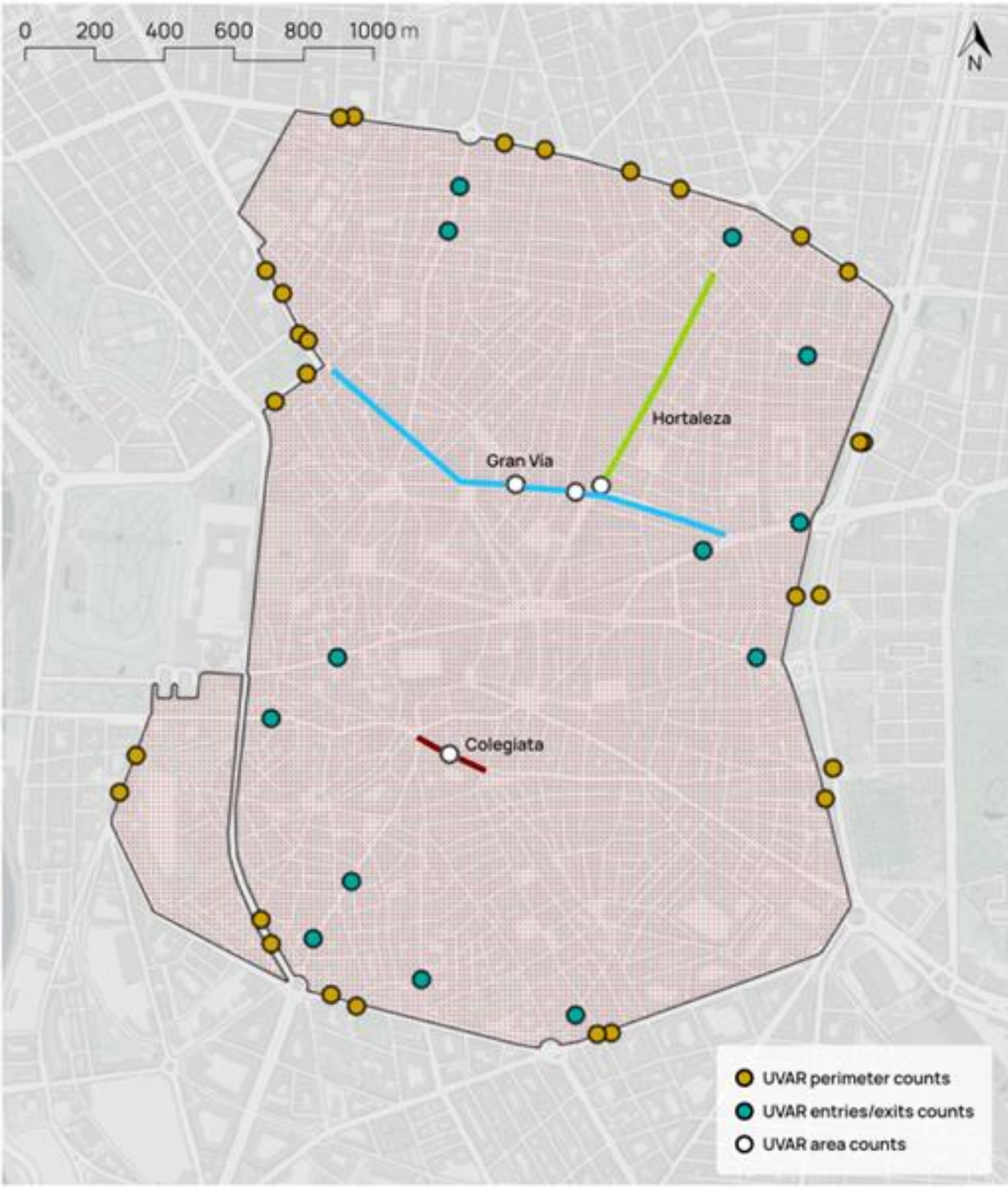


Figure 55. Traffic counts selected for the analysis of traffic evolution in the UVAR zone

		2017												2018												2019												2020	
MAD_CENTRAL	NOM_CORR	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2
BORDER	Alberto Aguilera E-O	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Alberto Aguilera O-E	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Carranza E-O	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Carranza O-E	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Genova E-O	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Genova O-E	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Gran Via San Francisco N-S	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Gran Via San Francisco S-N	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Prado Cibeles N-S	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Prado Cibeles S-N	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Prado Neptuno N-S	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Prado Neptuno S-N	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Princesa N-S	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Princesa S-N	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Puerta Toledo E-O	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Puerta Toledo O-E	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Pz España Cuesta N-S	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Pz España Cuesta S-N	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Pz España Princesa N-S	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Pz España Princesa S-N	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Recoletos N-S	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Recoletos S-N	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Ronda Segovia N-S	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Ronda Segovia S-N	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Ronda Valencia E-O	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Ronda Valencia O-E	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Sagasta E-O	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BORDER	Sagasta O-E	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
IN	Alcala Cedaceros	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
IN	Alcala IN	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
IN	Cuesta San Jeronimo OUT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
IN	Mayor IN	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
IN	San Bernardo IN	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
IN	Toledo IN	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
OUT	Barbara Braganza OUT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
OUT	Miguel Servet OUT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
OUT	Ribera Curtidores OUT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
OUT	San Bernardo OUT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
OUT	Santa Barbara OUT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
OUT	Segovia OUT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
OUT	Toledo OUT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Figure 56. Data availability in the selected counts for the period January 2017 - February 2020. Red cells correspond to months (columns) without valid data for a given count (rows)

There are few traffic counts within the UVAR zone that have enough data continuity to build up a spatially representative average of the traffic volumes. Many street works have been conducted in the area after the UVAR implementation (e.g. Atocha, Mayor, Gran Vía), and the traffic volumes in some streets are heavily affected by the traffic diversions put in place after the start of Plaza España refurbishment in summer 2019 (e.g. Santo Domingo). This limits the capacity to distinguish between reduction in through traffic (origin and destination outside of the UVAR zone) and connection traffic (origin or destination within the UVAR zone). From the data collected in the selected counts, the following monthly indicators are computed:

- Entry/exit traffic variation rate (cmn): percentage change in the total number of vehicles accessing or exiting the UVAR zone during a month m in 2017 and the same month n in 2018 or 2019. For each pair of months, only the traffic counts with ensured data quality in both months are used.

$$c_{mn} = \frac{\sum_i C_{in} - \sum_i C_{im}}{\sum_i C_{im}}; i \in traffic\ counts\ with\ valid\ data\ in\ m\ and\ n$$

- Perimeter traffic variation rate (pmn): percentage change in the total number of vehicles on the perimeter of the UVAR zone during a month m in 2017 and the same month n in 2018 or 2019. For each pair of months, only the traffic counts with ensured data quality in both months are used.

$$p_{mn} = \frac{\sum_i P_{in} - \sum_i P_{im}}{\sum_i P_{im}}; i \in traffic\ counts\ with\ valid\ data\ in\ m\ and\ n$$

According to these tailored indicators built from the traffic counts data, the traffic in the entries and exits to the UVAR zone decreased by almost 13% between 2017 and 2019. In line with the results from the aggregated indicator, part of the reduction can be attributed to the capacity restrictions conducted in 2018 (traffic fell 5,5% between 2017 and 2018), as can be seen in Table 9. However, the decrease in entry/exit traffic was more intense in 2019. The UVAR implementation was not followed by an increase in the traffic levels of the perimeter streets, since traffic volume actually dropped 1,4% along these streets in 2019 compared to 2018. This evolution is similar to the traffic variation in the whole city (-1,5%). The decrease in traffic volumes was particularly intense between January 2019 and July 2019 (down to -20% in the entries and exits of the UVAR zone), as can be seen in Figure 57.

Area	2018 vs 2017	2019 vs 2017	2019 vs 2018
UVAR perimeter	-2,9%	-4,8%	-1,4%
UVAR entries/exits	-5,5%	-12,8%	-8,1%
City total	-1,9%	-3,4%	-1,5%

Table 9. Yearly traffic variations in the UVAR zone compared to total in the city

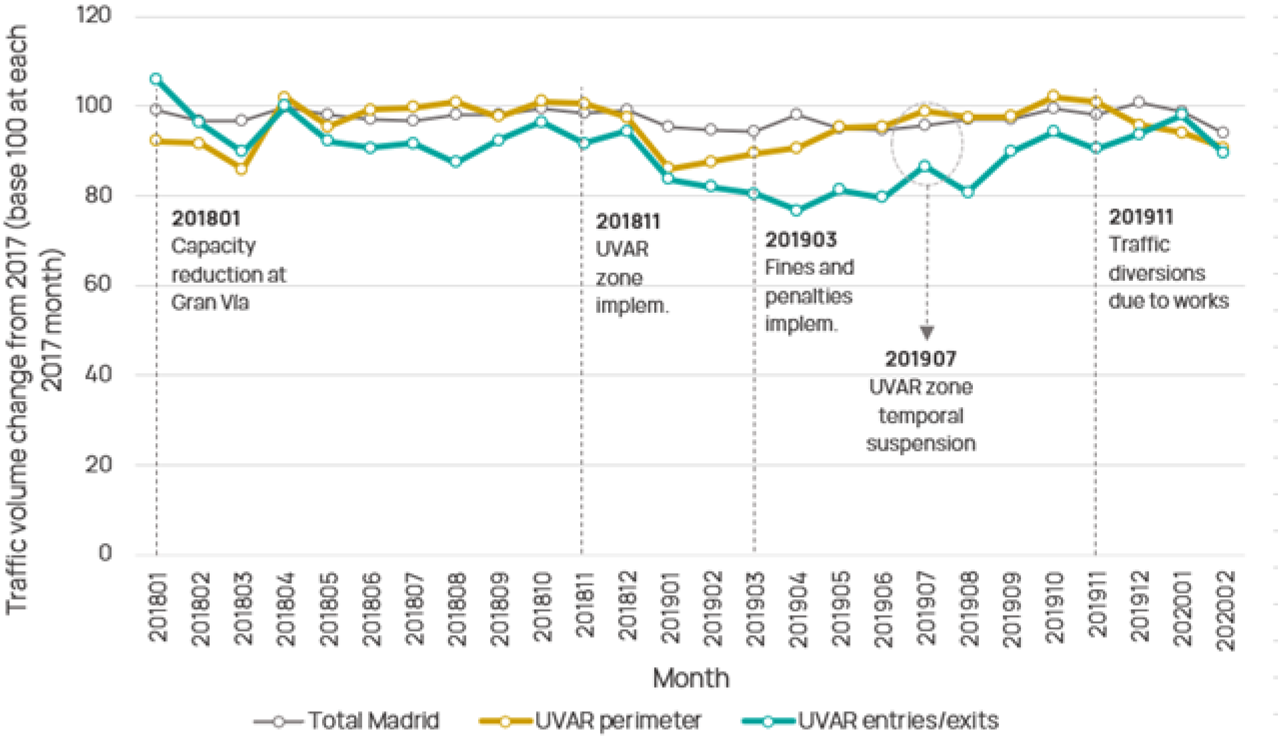


Figure 57. Traffic evolution from 2017: Madrid Central perimeter, in/out and within counts vs. total in the city

The data from some traffic counts within the UVAR zone suggests that at least part of this reduction was related to through traffic. The drop in main streets across the UVAR zone was higher than in secondary streets within the area, such as Colegiata street (Figure 58). However, some reduction can be seen also in these streets since the implementation of the UVAR zone, together with a reduction in the total traffic volumes in the city according to the aggregated indicator. Hence, it is likely that the reduction in traffic volumes is a result of a reduction both in through traffic and in connection traffic.

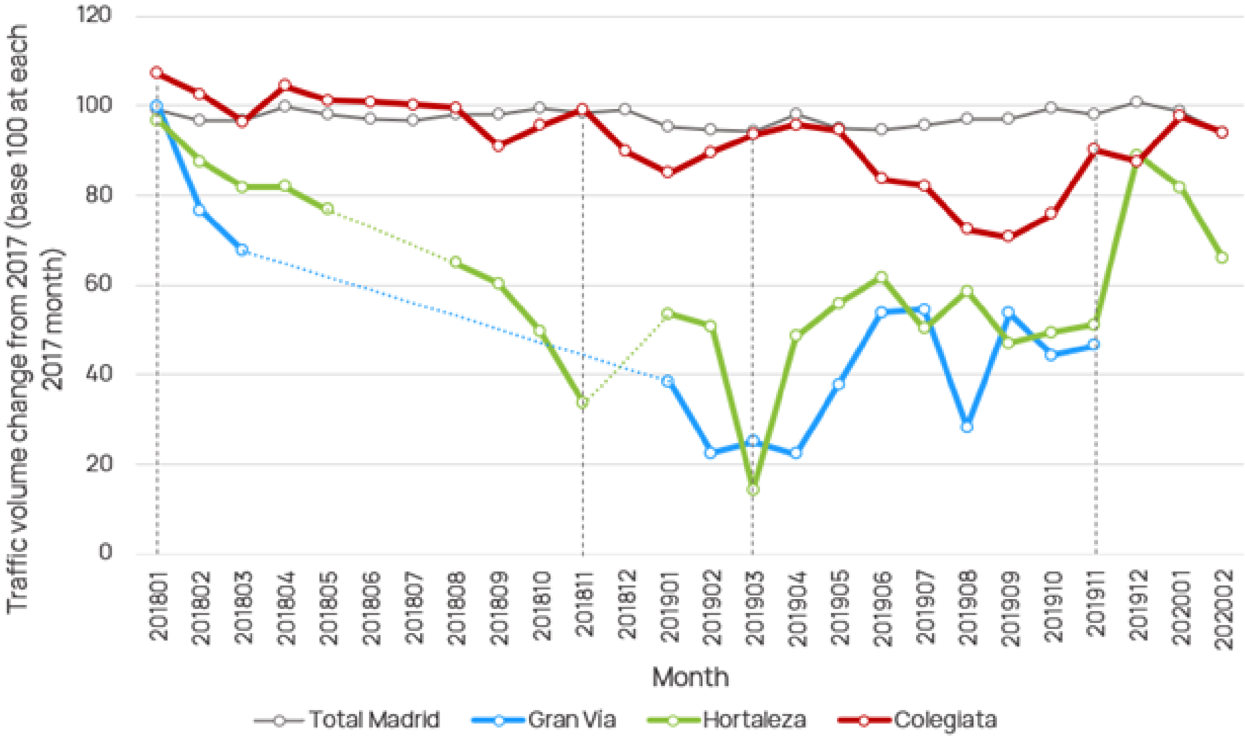


Figure 58. Traffic evolution from 2017: traffic counts within UVAR zone

3.1.4.3.3. Impact of the UVAR zone in overall mobility demand

Once some reduction in connection traffic is proved to be likely, the second experiment explores to what extent the following two hypothesis can explain this reduction:

Modal shift hypothesis: the reduction was caused by a modal shift from private vehicles to alternative modes. The eventual disutility perceived by some citizens due to the limited accessibility by private vehicle would have been not enough to prevent most of them from performing activities in the city centre.

Dissuaded demand hypothesis: the reduction was caused by an activity relocation effect due to the worse accessibility by private vehicle. The eventual disutility perceived by some citizens due to this limited accessibility would have led them to find alternative locations to perform their activities.

3.1.4.3.3.1. Overall mobility demand indicators

This is done by extracting overall mobility origin-destination matrices from mobile network data, which capture all trips performed in the Madrid region regardless of the transport mode, to evaluate the feasibility of the dissuaded demand hypothesis. The following concepts are defined for this experiment:

- Total trip variation rate (tmn): percentage change in the total number of trips within Madrid region between a period m (Tm) before and a period n (Tn) after the UVAR zone implementation. The metric is defined at region level to take into account that activity relocation effects may involve leisure and commercial attractions outside Madrid municipality, particularly for those residents in metropolitan municipalities.

$$t_{mn} = \frac{T_n - T_m}{T_m}$$

- External trip variation rate (e_{mn}): percentage change in the number of trips with origin and destination outside the UVAR zone between a period m (E_m) before and a period n (E_n) after the UVAR zone implementation.

$$e_{mn} = \frac{E_n - E_m}{E_m}$$

- Connection trip variation rate (c_{mn}): percentage change in the number of trips with origin or destination within the UVAR zone between a period m (C_m) before and a period n (C_n) after the UVAR zone implementation.

$$c_{mn} = \frac{C_n - C_m}{C_m}$$

- Internal trip variation rate (u_{mn}): percentage change in the number of trips with origin and destination within the UVAR zone between a period m (U_m) before and a period n (U_n) after the UVAR zone implementation.

$$u_{mn} = \frac{U_n - U_m}{U_m}$$

- Connection differential indicator (d_{mn}): difference between connection trip variation rate and total trip variation rate, both calculated between a period m before and a period n after the UVAR zone implementation. This is multiplied by 100 to be expressed in percentage points.

$$d_{mn} = 100 \cdot (c_{mn} - t_{mn})$$

3.1.4.3.3.2. Evolution of the indicators

In a first stage, the value of d_{mn} for the different post periods n is analysed. Negative values of d_{mn} would support the dissuaded demand hypothesis, since connection demand would have grown less or decrease more than total demand. A limitation of this approach is that there may be some seasonal effects in the differences between t_{mn}, e_{mn} and c_{mn} rates mixing with the potential effects of the UVAR zone. This is alleviated by including October 2019 as one of the post periods, fully comparable to October 2018.

Overall mobility demand was generally lower in 2019 than in 2018 in the Madrid region (Table 10). All 2019 months analysed recorded lower values than October 2018, including October 2019 (-5.0%). In general, the number of trips with origin or destination in the UVAR zone experienced more decrease, with the exception of October 2019. Among the categories that segment total demand according to their relation with the UVAR zone, internal mobility to the UVAR zone were generally the one that experienced more decrease.

Indicator	February 2019	April 2019	June 2019	July 2019	October 2019
Total trip variation rate (t_{mn})	-7,4%	-6,9%	-4,5%	-13,5%	-5,0%
External trip variation rate (e_{mn})	-7,2%	-6,5%	-4,0%	-12,9%	-5,0%
Connection trip variation rate (c_{mn})	-9,1%	-10,9%	-10,3%	-20,9%	-4,0%
Internal trip variation rate (u_{mn})	-9,9%	-14,0%	-9,1%	-21,8%	-6,4%
Connection differential indicator (d_{mn})	-1,71	-3,97	-5,88	-7,35	1,02

Table 10. Evolution of overall mobility demand indicators from October 2018 (monthly daily averages)

There were no specific patterns in the differences of the variation of overall mobility depending on the weekday, as Table 11. Evolution of connection differential indicator for overall mobility demand from October 2018 (by weekday)Table 11 shows. Except in February 2019, Mondays and Tuesdays show the lowest difference between UVAR zone reduction and the mobility in the whole Madrid region, while the day with higher difference was Saturday in three out of five months.

Connection differential indicator (d_{mn})	February 2019	April 2019	June 2019	July 2019	October 2019
Monday	-2,50	-3,00	-3,69	-4,72	2,37
Tuesday	-3,19	-3,70	-5,36	-5,37	1,55
Wednesday	-3,19	-4,83	-5,63	-6,36	0,69
Thursday	-2,53	-4,66	-6,99	-7,39	0,72
Friday	-2,43	-2,93	-7,23	-7,14	1,44
Saturday	-1,14	-5,71	-10,69	-9,92	1,51
Sunday	-0,44	-3,86	-9,56	-8,98	2,98

Table 11. Evolution of connection differential indicator for overall mobility demand from October 2018 (by weekday)

Even though the negative trend on the connection differential indicator may be an evidence supporting the dissuaded demand hypothesis, it has to be taken into account that in the yearly comparison (October 2019 vs. October 2018) no specific reduction in the connection trip volumes was detected in relation to the total trip volumes (d_{mn}=1.02). The negative trend in the previous months may indicate either (1) seasonal effects causing a reduction of these connection trips, (2) a transitory dissuaded demand effect until travellers find alternatives to private vehicle use or (3) a combination of both.

3.1.4.3.3.3. Relation of the indicators with private vehicle modal shares

The use of modal share data from the 2018 mobility household survey can provide further evidence supporting or rejecting the dissuaded demand hypothesis. This is done by comparing the connection differential indicator values and private vehicle modal shares at two levels: (1) OD-pair and (2) period of the day.

A piece of evidence supporting the dissuaded demand hypothesis would be that a particular decrease of the number of trips to/from the UVAR zone is observed in those origin-destination pairs that scored a higher private vehicle modal share in the survey. Hence, the above metrics will be particularised for the 21 districts in the Madrid municipality (excluding Centro district, which is extensively covered by the UVAR zone) and the 25 municipalities of the region with higher overall demand volumes to/from the UVAR zone according to the household survey (>1,000 daily trips). The total and connection trip rates for a specific area i are defined as follows:

- **Total trip variation rate for an area i** (t_{mni}): percentage change in the total number of trips between an area i and the whole region of Madrid, and vv., between a period m (T_m) before and a period n (T_n) after the UVAR zone implementation. It includes trips with origin and destination within the area i.
- **Connection trip variation rate for an area i** (c_{mni}): percentage change in the number of trips between an area i and the UVAR zone, and vv., between a period m (C_{mi}) before and a period n (C_{ni}) after the UVAR zone implementation.

The same logic can be applied to the period of the day, by checking if the potential decrease in the number of trips was higher at the hours of the day when the private vehicle modal share is larger. The total and connection trip rates for a specific hourly period p are defined as follows:

- **Total trip variation rate for an hour p** (t_{mnp}): percentage change in the total number of trips within Madrid region, during an hour of the day p, between a period m (T_{mp}) before and a period n (T_{np}) after the UVAR zone implementation.

- **Connection trip variation rate for an area i** (c_{mnp}): percentage change in the number of trips with origin or destination within the UVAR zone, during an hour of the day p, between a period m (C_{mp}) before and a period n (C_{np}) after the UVAR zone implementation.

The potential relation between the particularised connection differential indicators (d_{mni} and d_{mnp}) and the modal share of private vehicle in each demand segment i and p is explored to find evidence supporting the hypothesis raised.

When analysing OD pairs between the UVAR zone and the rest of Madrid city districts, some relation between the private vehicle modal share in these OD pairs and the overall mobility variation seems to emerge. Table 12 shows the coefficient of determination (R^2) values for the linear regressions of the connection differential indicator compared to private vehicle modal shares and Figure 59 illustrates the October 2019 case. As can be seen, the singular behaviour of Barajas district distorts the regression, since the apparent relation turns to be much weaker if Barajas district is removed. Without it, R^2 values peak in June 2019 at 0.23, but remain near or below 0.10 the rest of analysed months. It has to be noted that the fitness of both distributions improves if only private cars are taken into account to calculate the modal share (excluding motorbikes) and improves even more if only driver trips are taken into account.

If the analysis is extended to metropolitan municipalities, both variables (connection differential indicator and private vehicle modal share in the flows to/from the UVAR zone) show no relation at all, as can be seen in Table 13 and Figure 60. The behaviour of very different OD pairs in terms of private car use (e.g. Boadilla del Monte with >50% modal share vs. Alcorcón <20% modal share) is very similar in terms of connection differential indicator.

Indicator	February 2019	April 2019	June 2019	July 2019	October 2019
Private vehicle modal share	0.31 (0.06)	0.30 (0.06)	0.43 (0.23)	0.25 (0.03)	0.25 (0.11)
Private car modal share	0.33 (0.10)	0.32 (0.08)	0.44 (0.25)	0.28 (0.05)	0.29 (0.15)
Private car (as a driver) modal share	0.44 (0.10)	0.43 (0.10)	0.52 (0.26)	0.34 (0.04)	0.30 (0.12)

Table 12. Coefficients of determination R^2 between the connection differential indicator from October 2018 (average Tuesday to Thursday) in the OD pairs between the UVAR zone and the other districts of the city, and the modal share of private vehicle trips. The value in brackets refers to the regression excluding Barajas district.

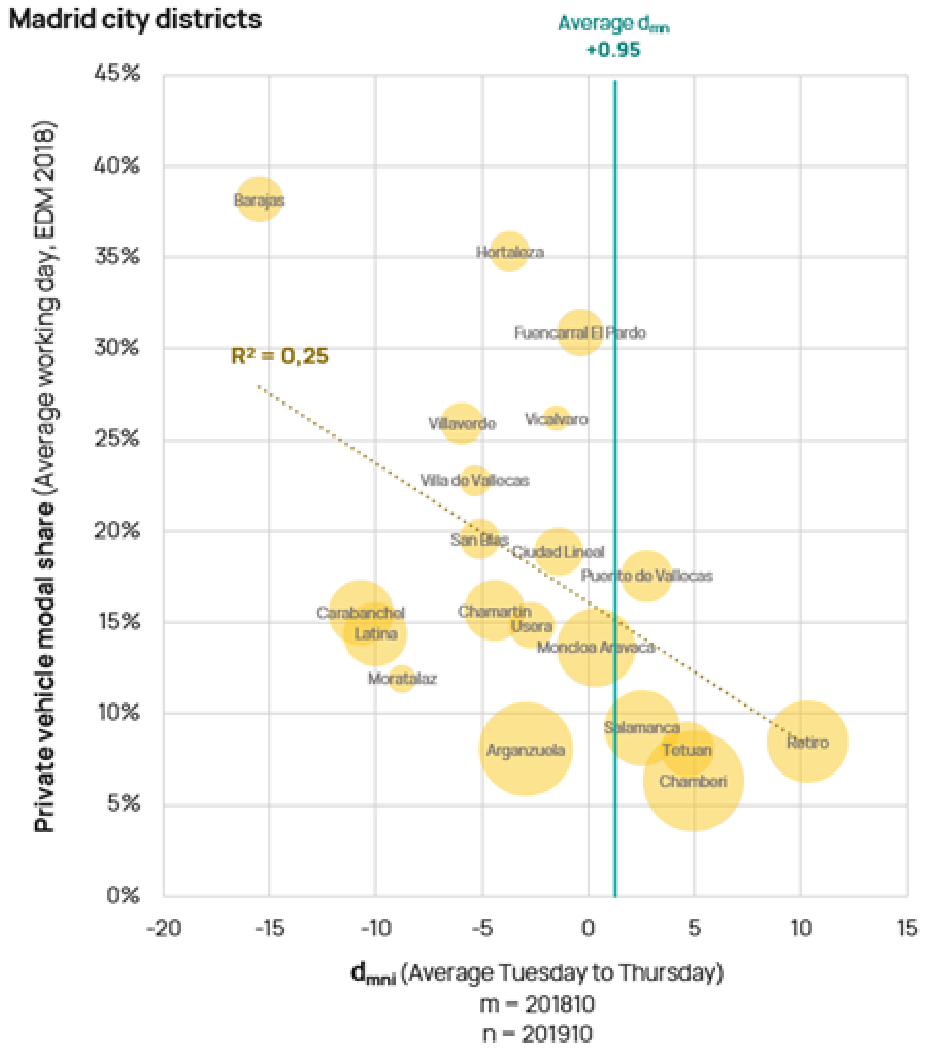


Figure 59. Connection differential indicator from October 2018 to October 2019 (average Tuesday to Thursday) in the OD pairs between the UVAR zone and the other districts of the city, compared to the modal share of private vehicle trips. The size of the points represents the number of connection trips in October 2018 in each OD pair.

Indicator	February 2019	April 2019	June 2019	July 2019	October 2019
Private vehicle modal share	0.04	0.09	0.01	0.04	0.03
Private car modal share	0.03	0.09	0.01	0.01	0.01
Private car (as a driver) modal share	0.05	0.02	0.00	0.01	0.01

Table 13. Coefficients of determination R^2 between the connection differential indicator from October 2018 (average Tuesday to Thursday) in the OD pairs between the UVAR zone and metropolitan municipalities, and the modal share of private vehicle trips. The value in brackets refers to the regression excluding Barajas district.

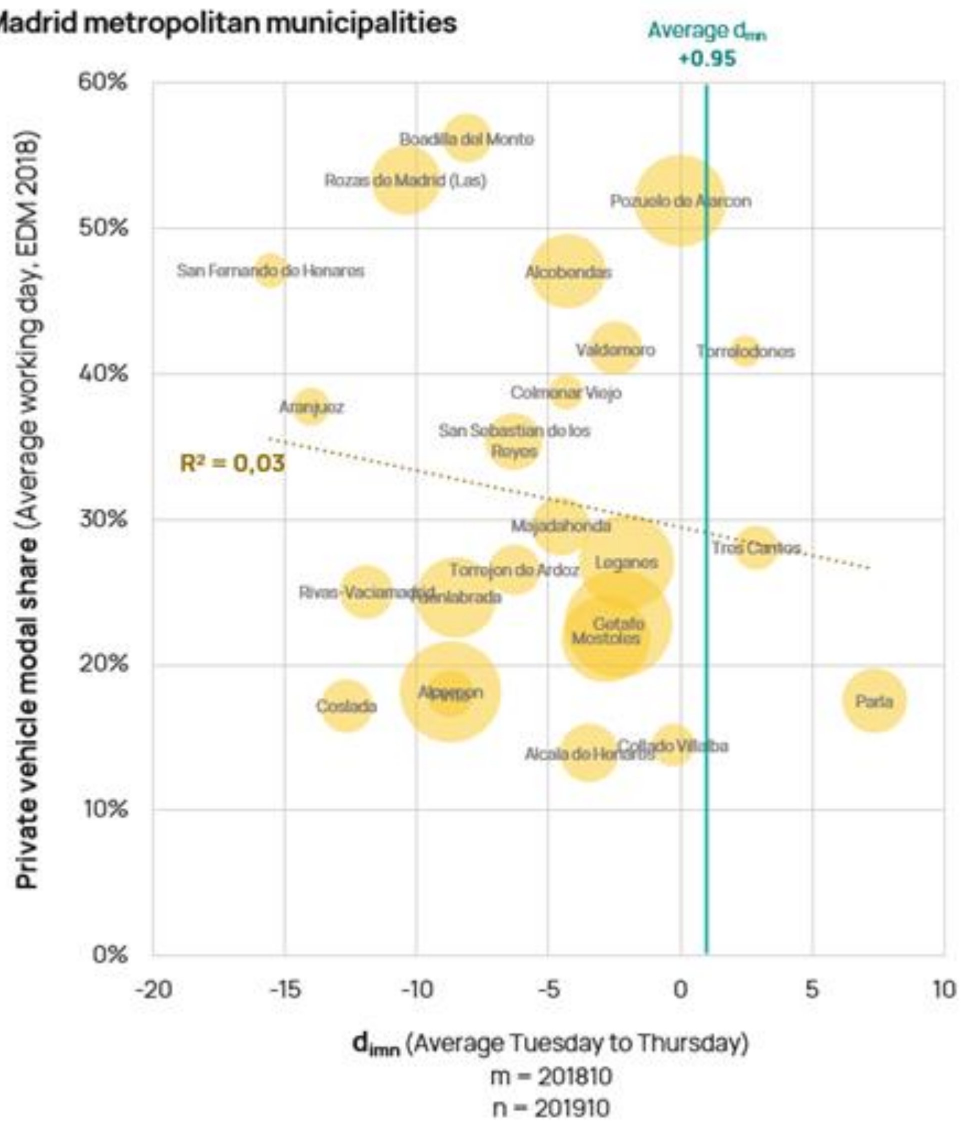


Figure 60. Connection differential indicator from October 2018 to October 2019 (average Tuesday to Thursday) in the OD pairs between the UVAR zone and metropolitan municipalities, compared to the modal share of private vehicle trips. The size of the points represents the number of connection trips in October 2018 in each OD pair.

A similar analysis can be done for the period of the day when the trip was originated. According to the survey, the modal share of trips made by private vehicle to the area designated as UVAR zone was higher at late evenings and at early mornings (Figure 61). Flows from UVAR zone to the rest of the city and from UVAR zone to the rest of the Madrid region follow a similar pattern. These intraday differences in modal share can be useful to explore if there is some relation with the variation of overall mobility demand.

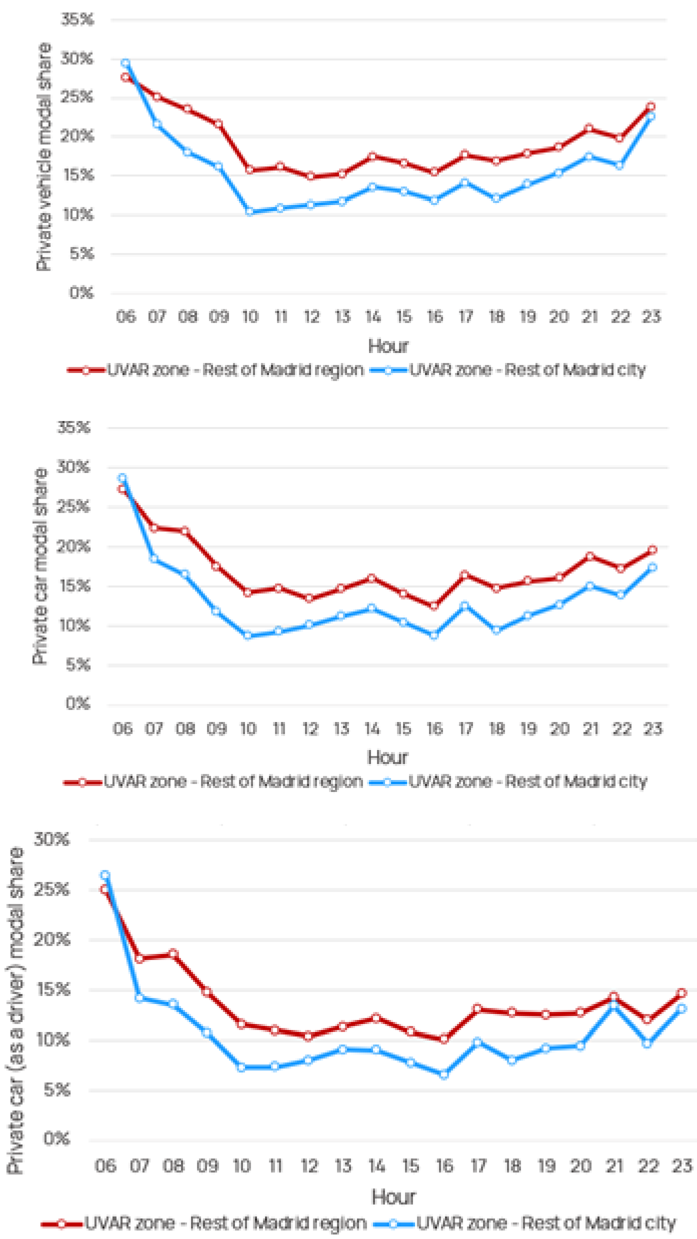


Figure 61. Private vehicle modal share (above, all private vehicle car trips; centre, private car trips; below, driving a private car) depending on the period of the day for trips with destination in the UVAR zone. Source: 2018 Household Survey.

The results show that there is no relation between the value of the connection differential indicator and the private vehicle modal share in the different times of the day at hourly level. Table 14 shows the R^2 values for the linear regressions of the connection differential indicator compared to private vehicle modal shares, together with the sign of the relation (positive indicates direct relation and negative indicates inverse relation). Figure 62 illustrates this for October 2019. The strongest relations are obtained for October 2019, but they show the opposite trend to the one that would support the dissuaded demand hypothesis: the overall mobility demand to/from the UVAR zone grew more than the regional overall mobility demand precisely in those hourly periods where the modal share of private vehicles is higher among the trips to/from the UVAR zone. This suggests that the activity relocation effect due to worse accessibility by private vehicle has been limited. In addition, the low fitness of the regressions for the previous months make unlikely the possibility of a transitory dissuaded demand effect until travellers find alternatives. Following this, it seems that the negative values of the connection differential indicator until October 2019 are more related to seasonal patterns. The absence of a strong relationship between OD pair level and period level private vehicle modal shares and the evolution of connection trips, together with the positive value of d_{mn} for October 2019, suggests that dissuaded demand is not the unique driver for the connection traffic reduction and that some modal shifts from private vehicles to alternative modes would have occurred.

Indicator	February 2019	April 2019	June 2019	July 2019	October 2019
Private vehicle modal share	0.00 (-)	0.09 (-)	0.02 (-)	0.02 (+)	0.18 (+)
Private car modal share	0.00 (+)	0.02 (-)	0.00 (-)	0.06 (+)	0.33 (+)
Private car (as a driver) modal share	0.00 (+)	0.00 (-)	0.02 (+)	0.13 (+)	0.32 (+)

Table 14. Coefficients of determination R^2 between the connection differential indicator from October 2018 (average Tuesday to Thursday) in each hour of the day, and the modal share of private vehicle trips.

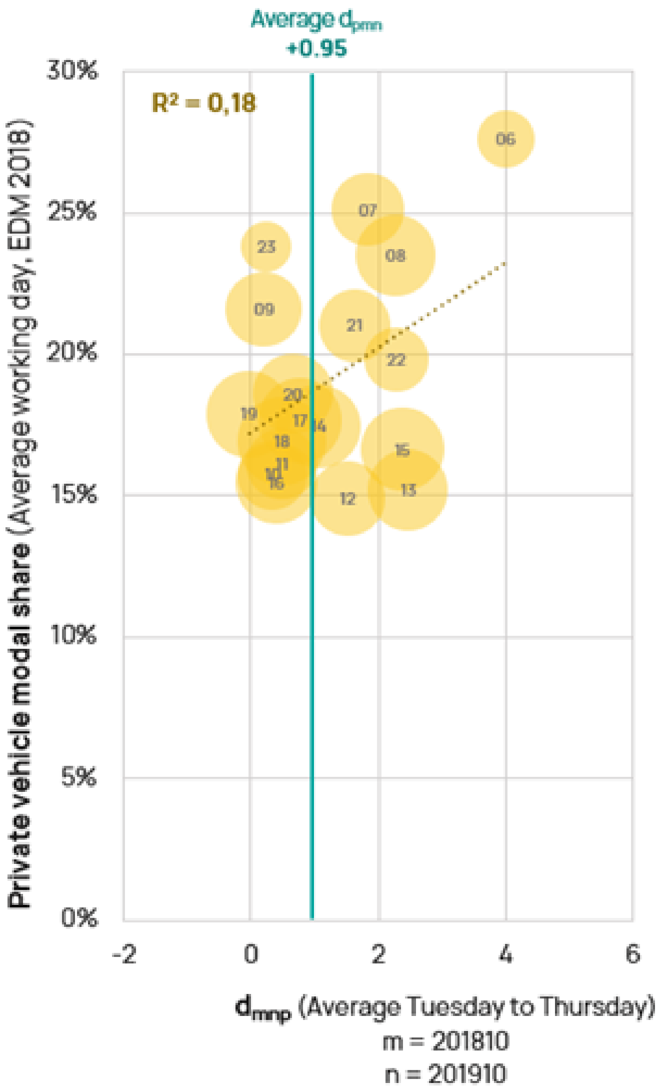


Figure 62. Connection differential indicator from October 2018 to October 2019 (average Tuesday to Thursday) in each hour of the day, compared to the modal share of private vehicle trips. The size of the points represents the number of connection trips in October 2018 in each hour of the day.

3.1.4.3.4. Characterisation of changes in shared mobility demand

Once it has been proved that modal shifts may have happened due to the implementation of the UVAR zone, it is possible to inquire into shared mobility operation data to check if these services captured part of the private vehicle demand. This is done by extracting shared mobility origin-destination matrices from the operation data. In an analogous process to the one described for overall mobility, if the trip volumes to/from the UVAR zone would have grown after the measure in a more intense way than the service whole demand this could indicate an attraction of some private car users to perform these connection trips. The concepts used for this analysis are analogous to the ones used for the overall demand evolution. The total rates are referred to the service areas of bike sharing and moto sharing systems.

3.1.4.3.4.1. Adaptation of overall mobility indicators to shared mobility demand

A limitation of this experiment is that the connection differential indicator is sensitive to the frequent changes in the service that shared mobility services have experienced in the first years of operation. While in the case of the analysed moto sharing service no major changes in the service area have been noticed from October 2018 to October 2019, bike sharing service supply increased significantly between these two periods, as shown in Figure 63. For instance, if the new docks were installed only in the areas external to the UVAR zone and no other changes have impact on the demand, the rate e_{mn} should be expected to grow faster than the rates c_{mn} and u_{mn} regardless of the impact of the UVAR zone implementation.

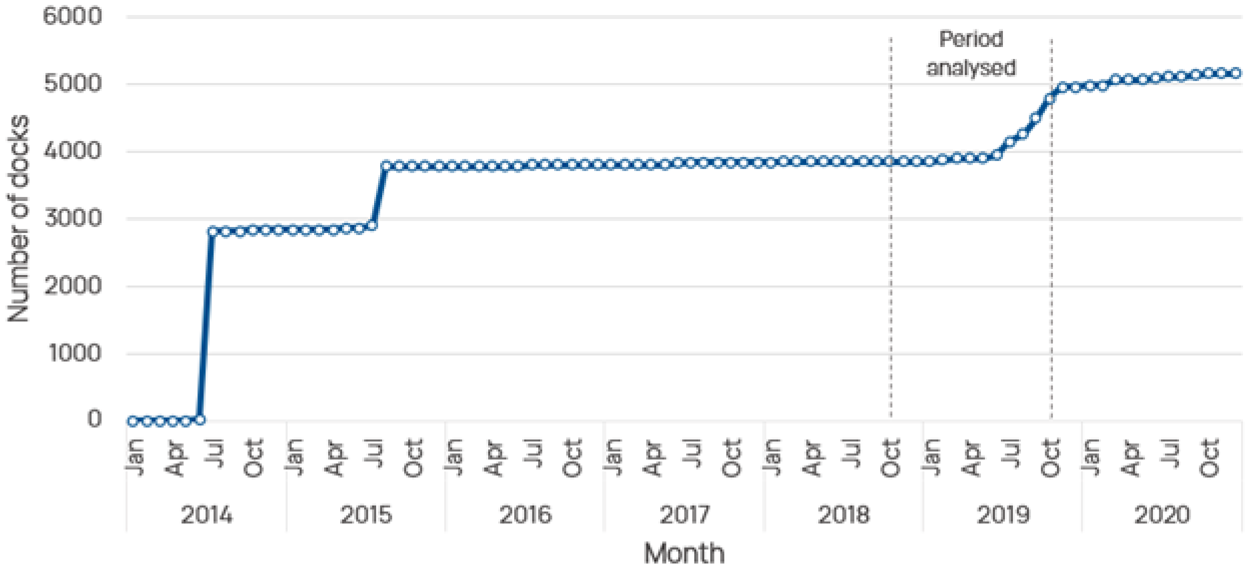


Figure 63. Monthly evolution of the number of available docks in the bike sharing service

The aforementioned indicators can be modified to reduce their sensitivity to the changes in supply. Instead of directly calculating the variation rates with the trips registered in periods m and n, the value of period n is corrected to assume that part of the increase in demand would have been caused by the increase in supply. The differential effect in variation rates is assumed to be sublinear with the increase in the number of dock pairs available for performing each type of trip. This assumption enables a first approach but could be further discussed as it has a direct impact on the results (e.g. Kabra et al (2020) found a similar sublinear relation between demand and walking distance to the nearest station). Consequently, the rates t_{mn} and c_{mn} which are used to compute the 'connection differential indicator' d_{mn} are transformed as follows:

$$t'_{mn} = \frac{(T_n - \frac{s_{tn} - s_{tm}}{s_{tn}} T_m) - T_m}{T_m}$$

$$c'_{mn} = \frac{(C_n - \frac{\sqrt{s_{en} s_{un}} - \sqrt{s_{em} s_{um}}}{\sqrt{s_{em} s_{um}}} C_m) - C_m}{C_m}$$

where:

- T_n and T_m are the total number of trips in the period m and in the period n respectively,
- C_n and C_m are the number of trips with origin or destination within the UVAR zone in the period m and in the period n respectively,
- s_{tn} and s_{tm} are the total number of docks available in the period m and in the period n respectively,
- s_{en} and s_{em} are the number of docks available outside the UVAR zone in the period m and in the period n respectively,
- s_{un} and s_{um} are the number of docks available within the UVAR zone in the period m and in the period n respectively.

For each service, the values of the connection trip variation rates c_{mn} and the connection differential indicator d_{mn} are monitored across the study periods to detect evidences supporting the role of the service as attractive for former users of private vehicles in their trips to/from the UVAR zone.

3.1.4.3.4.2. Evolution of the indicators

The two shared mobility services analysed had a fairly different evolution in 2019, as can be seen in Table 15 and Table 16. The demand of the bike sharing service increased 36.3% in October 2019 compared to October 2018, following an important increase in supply (28.6% more docks available in this period). Bike sharing demand has a strong seasonality component related to weather conditions. The demand of the moto sharing service decreased 41.5% in October 2019 compared to October 2018. This is likely to be due to the emergence of a strong competitor service in October 2018.

The evolution of connection trips in the bike sharing service was similar to the total demand until June 2019. Once the results are adjusted taking into account the added supply following the method explained in Section 3.2.3, it can be seen that the differences between connection trips and total trips in October 2019 are likely to be mostly related to the added supply outside the UVAR zone. The adjusted values of the connection differential indicator turned out to be bounded between -0.5 and +0.5 along the months analysed.

Indicator		February 2019	April 2019	June 2019	October 2019
Total trip variation rate (t_{mn})	No supply adjustment	-7,7%	-13,5%	37,9%	36,3%
	Supply adjustment	-8,7%	-14,5%	30,4%	7,7%
External trip variation rate (e_{mn})	No supply adjustment	-5,8%	-13,8%	44,3%	54,0%
	Supply adjustment	-5,8%	-13,9%	34,8%	15,2%
Connection trip variation rate (c_{mn})	No supply adjustment	-7,3%	-12,5%	37,1%	29,3%
	Supply adjustment	-8,9%	-14,1%	30,8%	7,3%
Internal trip variation rate (u_{mn})	No supply adjustment	-14,5%	-15,0%	19,8%	0,3%

	Supply adjustment	-17,6%	-18,2%	16,7%	-6,9%
Connection differential indicator (d_{mn})	**No supply adjustment**	**0,34**	**0,95**	**-0,80**	**-7,04**
	Supply adjustment	**-0,20**	**0,42**	**0,38**	**-0,42**
Number of docks in district Centro (1,245 in Oct. 2019)		+3,1%	+3,1%	+3,1%	+7,2%
Number of docks outside district Centro (2,613 in Oct. 2019)		+0,0%	+0,1%	+9,5%	+38,8%
Total number of docks (3,858 in Oct. 2019)		+1,0%	+1,1%	+7,5%	+28,6%

Table 15. Evolution of bike sharing service supply and demand indicators from October 2018 (monthly daily averages)

The evolution of connection trips in the moto sharing service shows stronger differences with the evolution of the total trips in the service. The connection differential indicator was consistently positive along the months analysed, indicating that the decreases in the moto sharing service demand were lower in the OD pairs between the UVAR zone and the rest of the city than in the rest of OD pairs (Table 16).

Indicator	February 2019	April 2019	June 2019	July 2019	October 2019
Total trip variation rate (t_{mn})	-29,7%	-39,0%	-23,3%	-33,9%	-41,5%
External trip variation rate (e_{mn})	-31,9%	-42,2%	-26,6%	-35,9%	-44,2%
Connection trip variation rate (c_{mn})	-24,3%	-32,0%	-14,6%	-29,5%	-35,1%
Internal trip variation rate (u_{mn})	-19,6%	-16,7%	-12,9%	-21,0%	-25,3%
Connection differential indicator (d_{mn})	**5,37**	**6,98**	**8,73**	**4,39**	**6,36**

Table 16. Evolution of moto sharing service demand indicators from October 2018 (monthly daily averages)

As it is the case with overall mobility demand, there are no specific weekday patterns regarding the differential evolution of the number of trips to/from the UVAR zone compared to the total demand of both services. Table 17 shows the weekday average of the connection differential indicator values along the analysed months in the bike sharing service, once adjusted to take into account the supply variation. Table 18 shows the values for the moto sharing service.

Connection differential indicator (d_{mn}) adjusted by supply variation	February 2019	April 2019	June 2019	October 2019
Monday	0,92	1,54	1,37	-1,21
Tuesday	-0,52	0,12	0,58	-1,68
Wednesday	-0,67	-0,26	0,12	-0,89
Thursday	1,10	0,20	1,33	-0,61
Friday	-0,68	0,82	-2,09	-0,41
Saturday	-1,58	-1,46	-2,35	1,45
Sunday	-2,07	0,46	-0,90	1,78

Table 17. Evolution of connection differential indicator for bike sharing demand from October 2018 (by weekday)

Connection differential indicator (d_{mn})	February 2019	April 2019	June 2019	July 2019	October 2019
Monday	5,26	9,91	8,22	7,10	6,37
Tuesday	10,99	12,69	21,44	14,00	8,77
Wednesday	3,61	4,81	11,19	8,10	5,66
Thursday	5,49	6,86	8,84	2,85	3,81
Friday	-0,44	7,37	3,14	4,96	9,05
Saturday	2,42	-2,08	-3,61	-0,39	1,17
Sunday	-1,04	1,36	-4,65	0,83	1,34

Table 18. Evolution of connection differential indicator for moto sharing demand from October 2018 (by weekday)

3.1.4.3.4.3. Relation of the indicators with private vehicle modal shares

The comparison of the connection differential indicator and private vehicle modal shares at OD pair level for the bike sharing system shows no relation. This has been first checked for an average working day, to be consistent with the data from the household survey. Table 19 shows the R^2 values for the linear regressions of the connection differential indicator compared to private vehicle modal shares. The values for the different alternative scopes for determining the modal share (only car, only car as a driver) show very similar results. Figure 64 illustrates the absence of relation for October 2019. The fact that few points are available (only 7 district OD pairs) poses a limit to this bivariate analysis. If the indicators are computed with all the days rather than from a Tuesday to Thursday average, similar results are obtained (Table 20 and Figure 65).

Indicator	February 2019	April 2019	June 2019	October 2019
Private vehicle modal share	0.00 (-)	0.04 (-)	0.02 (+)	0.00 (+)
Private car modal share	0.00 (-)	0.01 (-)	0.00 (+)	0.00 (-)
Private car (as a driver) modal share	0.01 (-)	0.01 (-)	0.01 (+)	0.00 (+)

Table 19. Coefficients of determination R2 between the connection differential indicator for bike sharing demand from October 2018 (average Tuesday to Thursday) in the OD pairs between the UVAR zone and the other districts of the city, and the modal share of private vehicle trips (average working day).

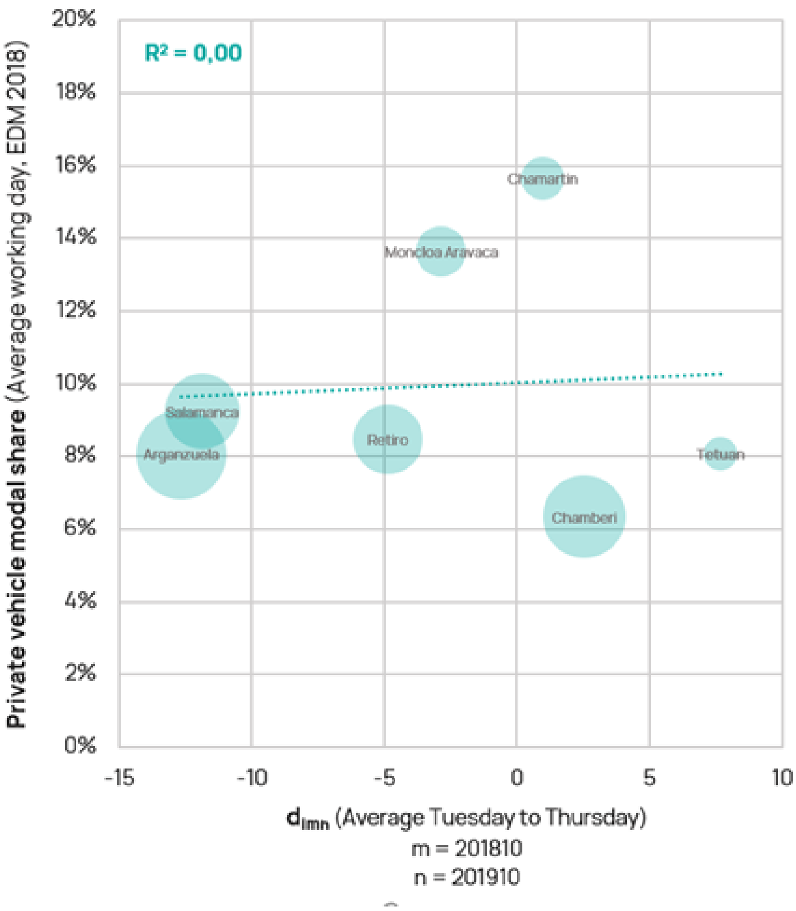


Figure 64 Connection differential indicator for bike sharing demand from October 2018 to October 2019 (average Tuesday to Thursday) in the OD pairs between the UVAR zone and the other districts of the city, compared to the modal share of private vehicle trips (average working day). The size of the points represents the number of connection bike sharing trips in October 2018 in each OD pair.

Indicator	February 2019	April 2019	June 2019	October 2019
Private vehicle modal share	0.00 (-)	0.01 (-)	0.02 (+)	0.00 (+)
Private car modal share	0.01 (-)	0.00 (-)	0.00 (+)	0.01 (-)
Private car (as a driver) modal share	0.01 (-)	0.00 (-)	0.00 (+)	0.00 (-)

Table 20. Coefficients of determination R2 between the connection differential indicator for bike sharing demand from October 2018 (all days) in the OD pairs between the UVAR zone and the other districts of the city, and the modal share of private vehicle trips (average working day).

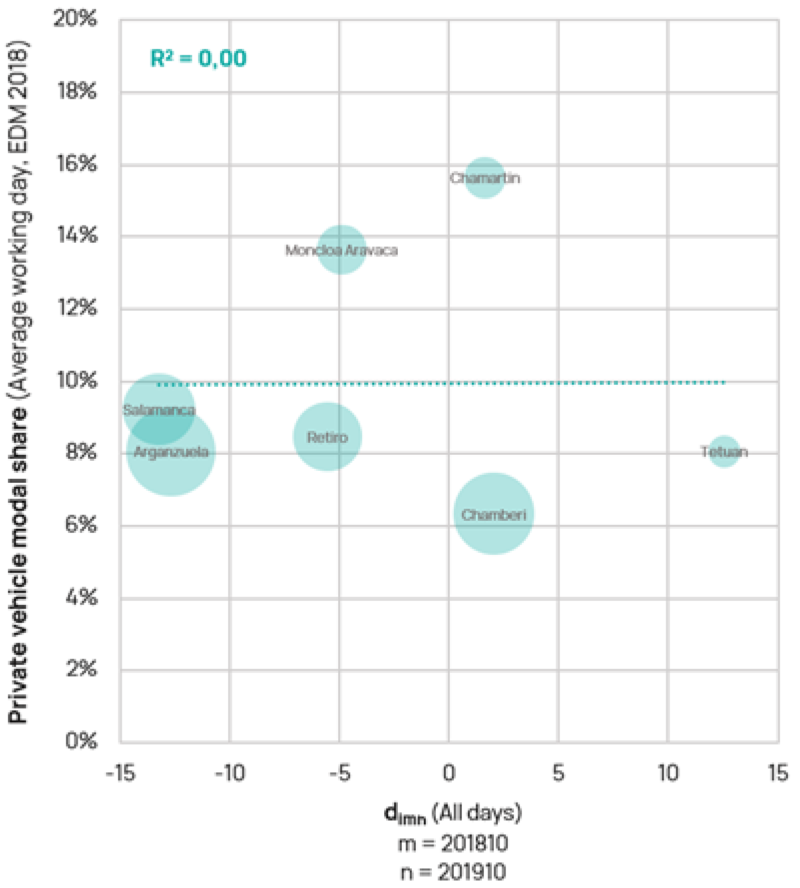


Figure 65. Connection differential indicator for bike sharing demand from October 2018 to October 2019 (all days) in the OD pairs between the UVAR zone and the other districts of the city, compared to the modal share of private vehicle trips (average working day). The size of the points represents the number of connection bike sharing trips in October 2018 in each OD pair.

In the case of moto sharing, the results show strong variations among the different months analysed. Until October 2019, the relation between the connection differential indicator and the private vehicle modal shares at OD pair level were slightly inverse, especially when the indicator is computed using working days (Table 21). R^2 values remain close to zero in the comparison with the indicator computed using all days until October 2019 (Table 22). This month, which enables a yearly comparison with October 2018 without seasonal influences, shows interesting results. While the comparison using working days keeps the relation inverse as in previous months (Figure 66), the comparison including weekends shows some relation (R^2=0.40), as can be seen in Figure 67. Fuencarral-El Pardo district, the one with highest private vehicle modal share among connection trips within the moto sharing service area, causes most of this shift. R^2 value drops to 0.27 if this district is removed from the comparison.

In order to check the robustness of this relation, a period level comparison was conducted to complement the OD pair level comparison. The results show similar patterns to the OD pair level comparison, with very weak relation in the case of working days indicator (Table 23 and Figure 68) and some positive relation if all days are considered (Table 24 and Figure 69). The R2 value for October 2019 is 0.21. The time period between 6am and 7am, which is the one with highest private vehicle modal share according to the survey, is the main driver of such relation, becoming much weaker if this time period is removed. Motorbike modal shares from the survey have been also compared, but no relation was found, except an inverse trend in the case of the hourly period comparison. It has to be taken into account that the sample size of the survey limits the number of motorbike trips captured.

Indicator	February 2019	April 2019	June 2019	July 2019	October 2019
Private vehicle modal share	0.24 (-)	0.22 (-)	0.18 (-)	0.07 (-)	0.06 (-)
Private car modal share	0.25 (-)	0.18 (-)	0.20 (-)	0.07 (-)	0.04 (-)
Private car (as a driver) modal share	0.28 (-)	0.20 (-)	0.30 (-)	0.05 (-)	0.06 (-)
Private motorbike modal share	0.00 (+)	0.22 (-)	0.01 (+)	0.01 (-)	0.19 (-)

Table 21. Coefficients of determination R2 between the connection differential indicator for moto sharing demand from October 2018 (average Tuesday to Thursday) in the OD pairs between the UVAR zone and the other districts of the city, and the modal share of private vehicle trips (average working day).

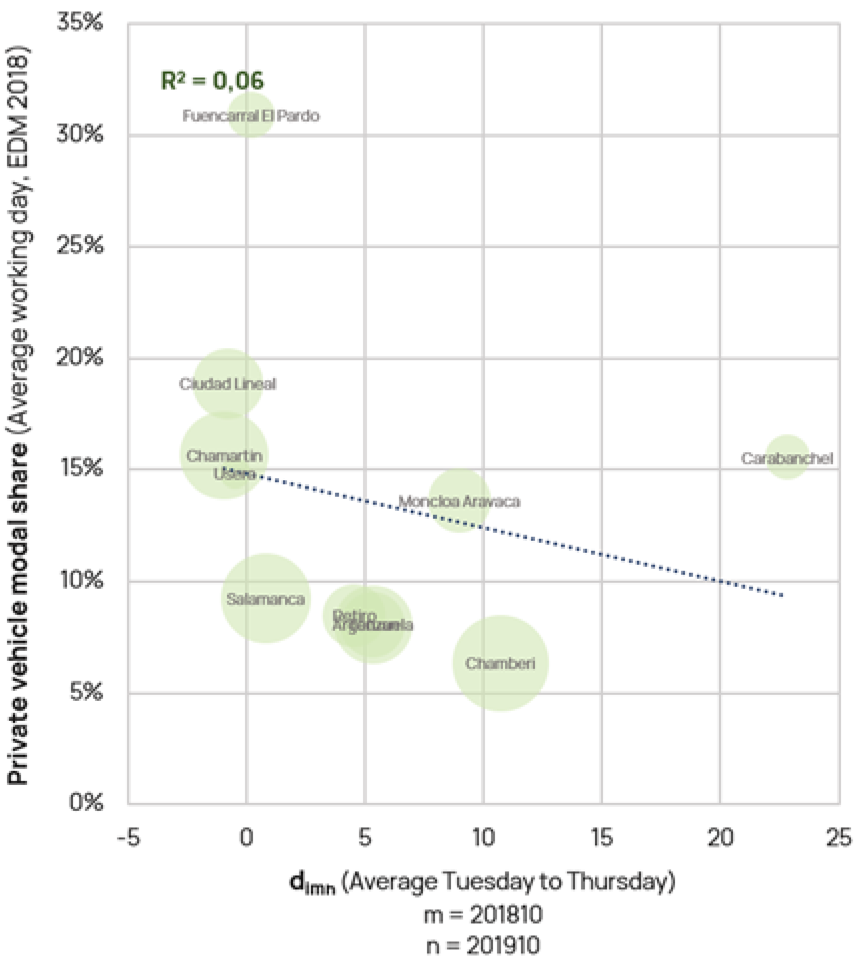


Figure 66. Connection differential indicator for moto sharing demand from October 2018 to October 2019 (average Tuesday to Thursday) in the OD pairs between the UVAR zone and the other districts of the city, compared to the modal share of private vehicle trips (average working day). The size of the points represents the number of connection moto sharing trips in October 2018 in each OD pair.

Indicator	February 2019	April 2019	June 2019	July 2019	October 2019
Private vehicle modal share	0.01 (+)	0.07 (-)	0.05 (+)	0.01 (+)	0.40 (+)
Private car modal share	0.01 (+)	0.06 (-)	0.03 (+)	0.01 (+)	0.44 (+)
Private car (as a driver) modal share	0.00 (+)	0.09 (-)	0.01 (+)	0.01 (+)	0.39 (+)
Private motorbike modal share	0.00 (+)	0.09 (-)	0.16 (+)	0.00 (+)	0.01 (-)

Table 22. Coefficients of determination R2 between the connection differential indicator for moto sharing demand from October 2018 (all days) in the OD pairs between the UVAR zone and the other districts of the city, and the modal share of private vehicle trips (average working day).

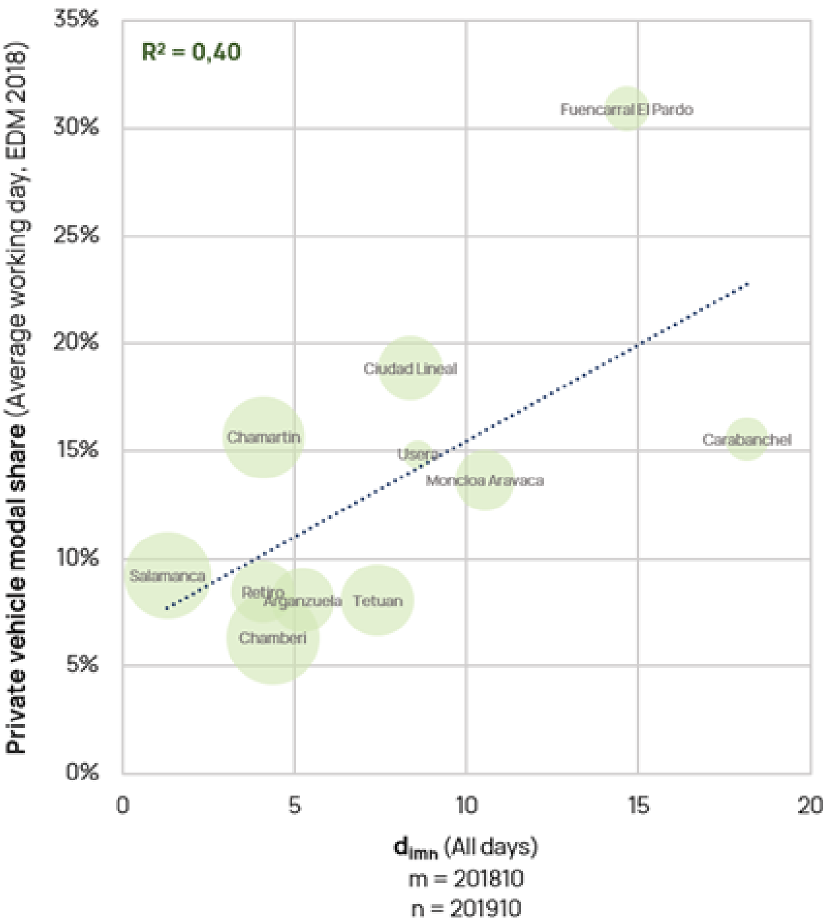


Figure 67. Connection differential indicator for moto sharing demand from October 2018 to October 2019 (all days) in the OD pairs between the UVAR zone and the other districts of the city, compared to the modal share of private vehicle trips (average working day). The size of the points represents the number of connection moto sharing trips in October 2018 in each OD pair.

Indicator	February 2019	April 2019	June 2019	July 2019	October 2019
Private vehicle modal share	0.03 (-)	0.03 (+)	0.03 (+)	0.09 (+)	0.05 (+)
Private car modal share	0.02 (-)	0.08 (+)	0.06 (+)	0.15 (+)	0.13 (+)
Private car (as a driver) modal share	0.02 (-)	0.10 (+)	0.06 (+)	0.20 (+)	0.14 (+)
Private motorbike modal share	0.01 (-)	0.25 (-)	0.12 (-)	0.14 (-)	0.25 (-)

Table 23. Coefficients of determination R2 between the connection differential indicator for moto sharing demand from October 2018 (all days) in each hour of the day, and the modal share of private vehicle trips (average working day).

Figure 68. Connection differential indicator for moto sharing demand from October 2018 to October 2019 (all days) in each hour of the day, compared to the modal share of private vehicle trips (average working day). The size of the points represents the number of connection moto sharing trips in October 2018 in hourly period.

Indicator	February 2019	April 2019	June 2019	July 2019	October 2019
Private vehicle modal share	0.13 (-)	0.16 (+)	0.11 (+)	0.21 (+)	0.21 (+)
Private car modal share	0.10 (-)	0.25 (+)	0.22 (+)	0.21 (+)	0.34 (+)
Private car (as a driver) modal share	0.07 (-)	0.22 (+)	0.25 (+)	0.36 (+)	0.36 (+)
Private motorbike modal share	0.04 (-)	0.13 (-)	0.26 (-)	0.15 (-)	0.22 (-)

Table 24. Coefficients of determination R2 between the connection differential indicator for moto sharing demand from October 2018 (all days) in each hour of the day, and the modal share of private vehicle trips (average working day).

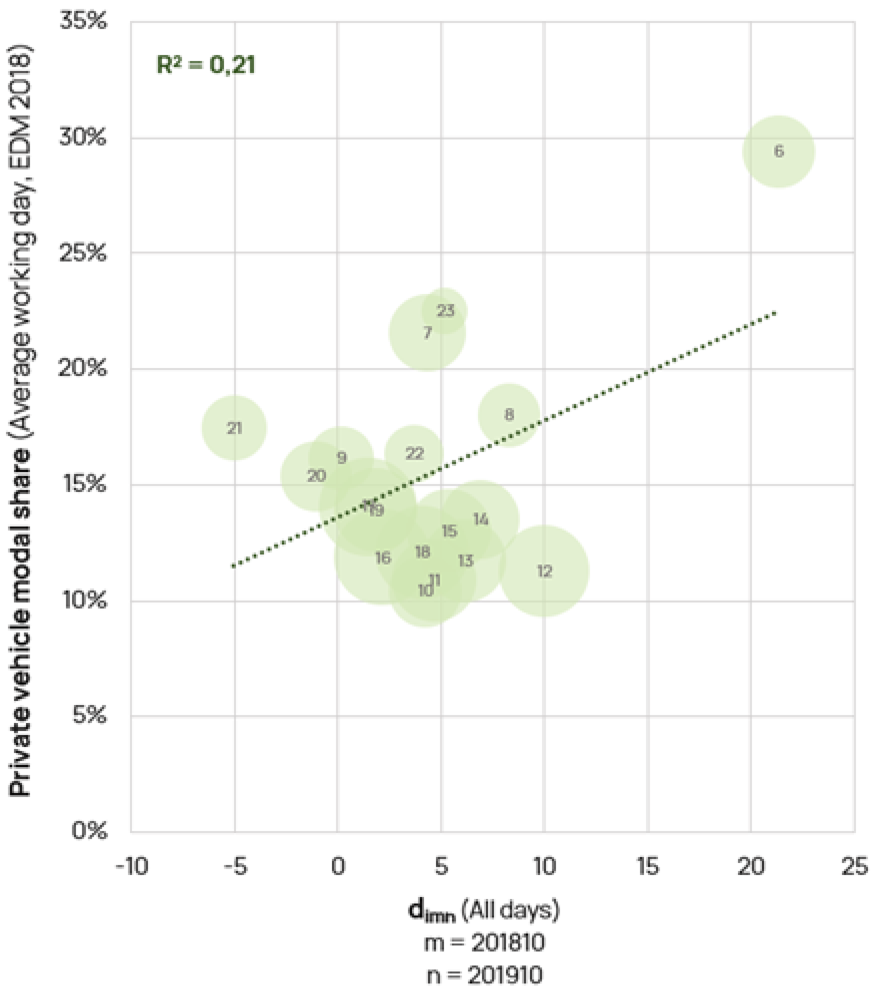


Figure 69. Connection differential indicator for moto sharing demand from October 2018 to October 2019 (all days) in each hour of the day, compared to the modal share of private vehicle trips (average working day). The size of the points represents the number of connection moto sharing trips in October 2018 in hourly period

3.1.4.3.5. Conclusions

The set of experiments conducted in this MOMENTUM case study draw interesting conclusions about the impact of emerging mobility solutions in Madrid and also about the technical opportunities and challenges that the use of big data sources entails for such assessments.

The implementation of the UVAR zone 'Madrid Central' caused a reduction in city centre traffic that is not fully reflected in the aggregated indicators built from permanent traffic counts, due to their spatial distribution. By processing the detailed data offered by more than 40 sensors in the entries, exits and the perimeter of the UVAR zone, the reduction in traffic was quantified. The fact that traffic fell more in those streets with higher propensity to through traffic indicates that most of the reduction was related to this type of flows. However, it seems that through traffic would not explain all the reduction, and that the traffic flows with origin or destination within the UVAR zone ('connection traffic') would have also fallen.

The use of mobile phone network data to obtain overall mobility OD matrices allows identifying whether the connection traffic reduction was a direct consequence of modal shifts to alternative modes, activity relocation to alternative places or a combination of both, given that these OD matrices include all the trips performed in the city regardless of the transport mode. While the months following the UVAR zone implementation were characterised by singular decreases in overall travel demand to/from the UVAR zone in comparison to the city total travel demand, this effect is no longer observable when comparing October 2019 and October 2018. Furthermore, the modal split data from the 2018 household survey was used to compare how the connection trip volumes behave in relation to the modal share that private vehicles achieved before the UVAR zone implementation, both at OD pair level and at hourly period level. The results reveal that the reduction in the number of trips to/from the UVAR zone was not particularly intense neither in the OD pairs nor in the hourly period with highest private vehicle modal share, with some exceptions (e.g. Barajas district). This suggests that the decreases observed between October 2018 and the first half of October 2019 would be related to seasonal patterns rather than to a possible transitory dissuaded demand effect until private vehicle users find attractive alternatives. According to these analyses, modal shifts from private vehicles to other modes would have occurred as a consequence of the implementation of the UVAR zone.

The availability of shared mobility operation data from the public bike sharing service and a moto sharing service allows to assess if these modes attracted some private car demand after the implementation of the UVAR zone. The evolution of the bike sharing demand does not offer evidence of this, since the growth in the OD pairs connecting the UVAR zone and the rest of the city was not especially intense. The lack of relation between these figures and the modal shares of private vehicles in each OD pair served by the bike sharing system suggests that it is unlikely that the service captured demand from private vehicles. This supports the previous findings in the literature (Ma et al., 2020). The evolution of moto sharing demand shows a less clear trend. In a context of large decreases in the service demand due to the emergence of other competitors, the flows to/from the UVAR zone behave better than the average, suggesting some gains in this demand segment counteracting the negative trend. The comparison of the evolution of the demand in these OD pairs with the pre-existing modal share of private vehicles does not clearly provide any evidence for a modal shift from cars. The demand indicators in working days, which are the ones fully consistent with the scope of the household survey, show no relation at OD pair level with the modal share of private vehicles. However, if the analysis is extended to all days including weekends, some relation emerges. This is also the case for the period level analysis. Given the novel character of moto sharing services there is scarce research about the ability of these services to capture car trips. The available studies are focused on adoption characterisation rather than in modal shift dynamics (Aguilera-Garcia et al., 2020), but the results of this case suggest that more research is needed in this direction.

With regard to the lessons learnt related to the application of big data sources and data fusion techniques to the assessment of emerging mobility solutions, the following aspects and limitations can be highlighted:

- Traffic counts are punctual measures that do not distinguish between through traffic and connection traffic, an issue that is particularly relevant for the assessment of UVAR zones. The data collected by the monitorization systems of these zones to impose fines and penalties could be used for this purpose.
- The rapid evolution of shared mobility systems, with constant changes in supply, make difficult to interpret the demand data they collect. In the context of this case study, bike sharing demand data was adjusted taking into account the increase in dock availability, but the correction applied was not supported by empirical evidence. A closer monitorization of shared mobility services would allow to better quantify the relation between supply (number of docks, number of vehicles, etc.) and demand.
- The fact that the mobility household survey was conducted only in working days limits the scope of the analysis. The findings about the moto sharing service, which suggest that some demand might have been captured during weekends in those OD pairs with highest private vehicle modal share according to the survey, have to be interpreted with caution because there is no modal share data covering weekends. This can be solved by the use of public transport smart card data.

3.1.5. Validation of the representative OD estimation

3.1.5.1. The Dataset

The data used for obtaining the OD matrices are the trips executed in June of 2019 in Madrid, Spain, having a total of 1259 zones. A total of 24 hourly OD matrices for each day of this month are created and therefore a total number of 30*24 OD matrices, one matrix for each day and each one of the 24 periods of the day. From the similarity analysis 5 groups were identified for June 2019:

- Group Saturdays: 1, 8, 15 and 29
- Group Sundays: 2, 9, 16, 22, 23 and 30
- Group Weekdays (mid-month): 3, 6, 11, 12, 13, 17, 18, 19 and 20
- Group Fridays: 7, 14, 21 and 28
- Group Weekdays (last-week): 24, 25, 26 and 27
- No clustered days: 4, 5 and 10

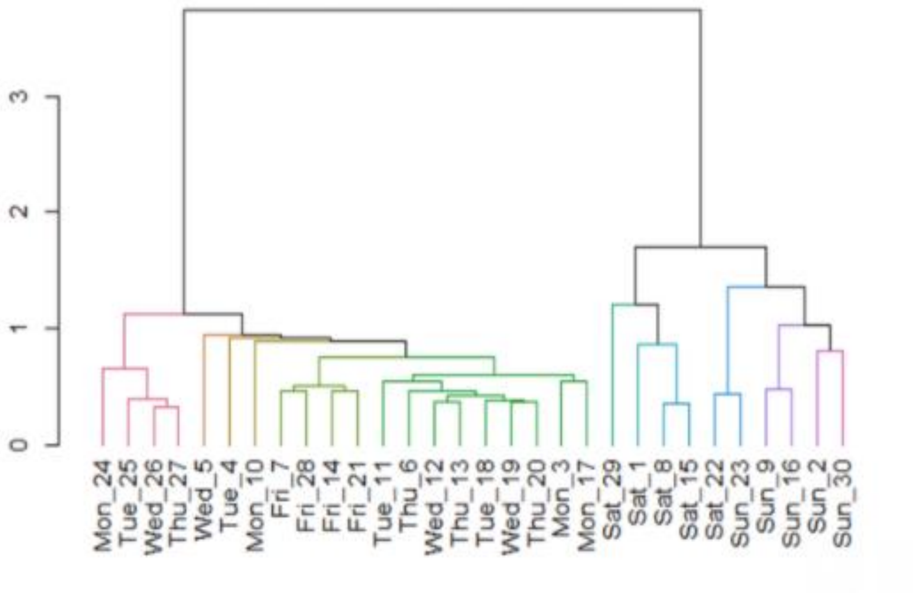


Figure 70. Dendrogram of the days clustering.

The pattern within the days of the month is presented in Figure 71, where weekdays and weekends are clearly identified. Still from the similarity analysis the days were clustered in more groups.

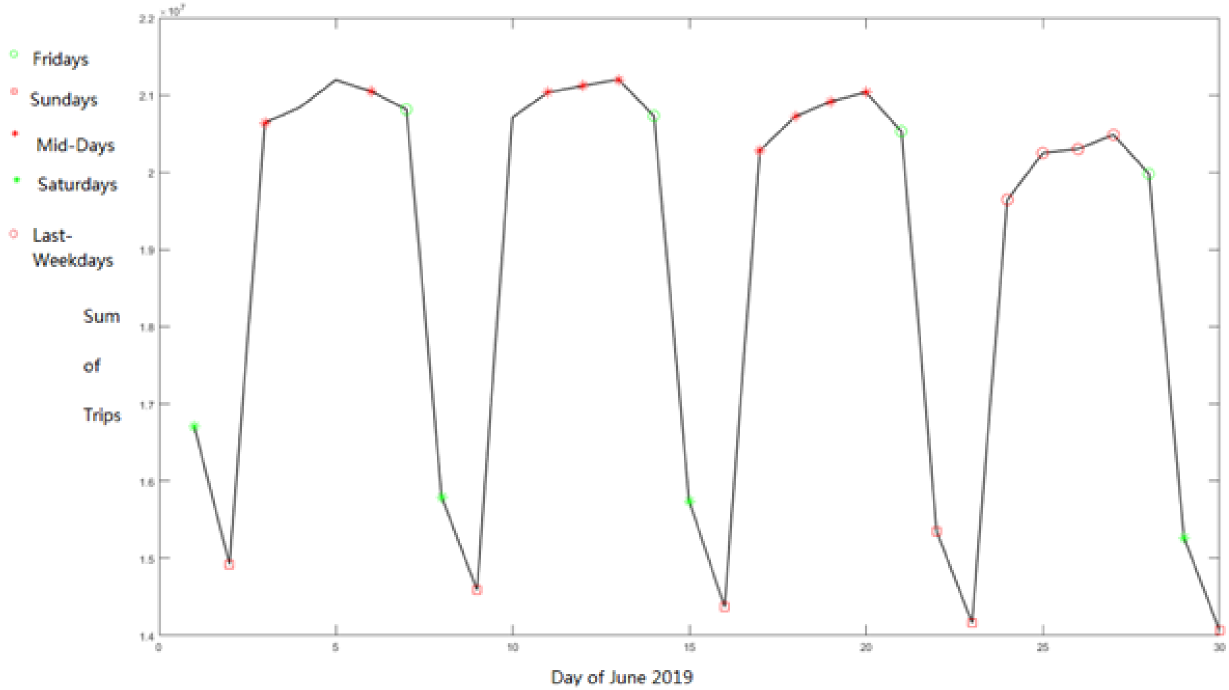


Figure 71. Total number of trips per day during June 2019

The focus of the analysis was on 3 periods that were considered as rush hours as the mean values of the average hourly trips reached local maxima, as can be seen in Figure 72. From Figure 72, we can see clearly that the three maxima that correspond to the three rush hours happen at 8:00-9:00, 14:00-15:00, and 19:00-20:00.

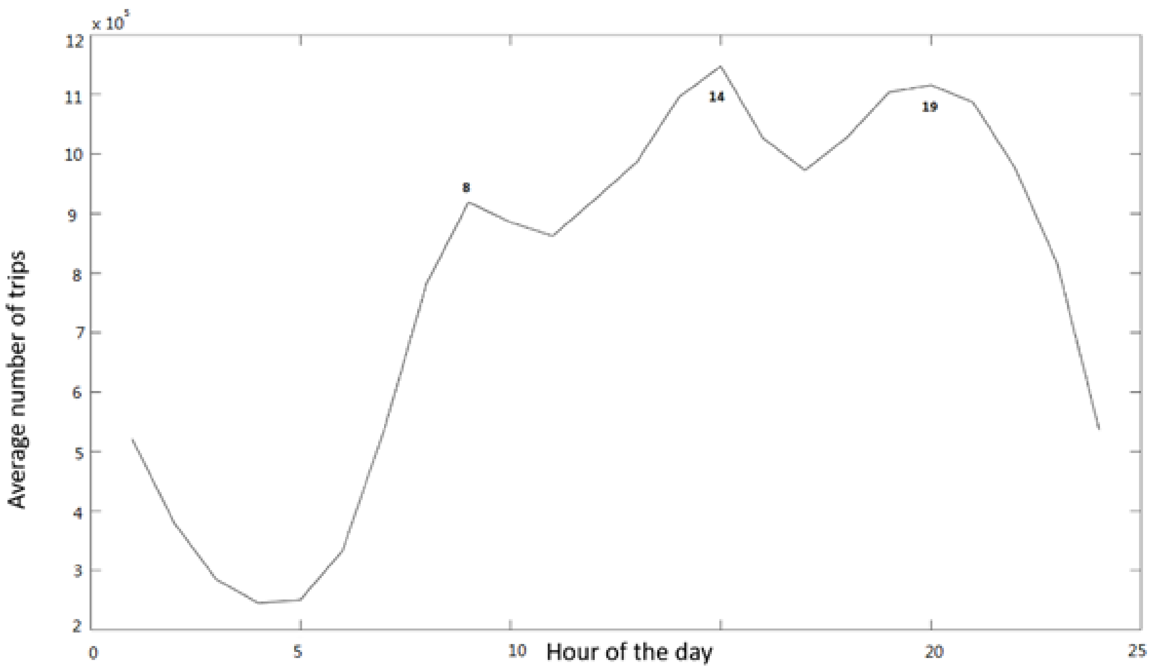


Figure 72. Hourly average number of trips during the day

Out of each group mentioned previously, the Fridays, the Mid-days of the week, the Last of the weekdays, the Saturdays and the Sundays, one day was chosen and the histograms for the 3 rush hours, periods 8, 14 and 19, are presented below. These figures are the distribution of the cell values in the columns of the OD matrix.

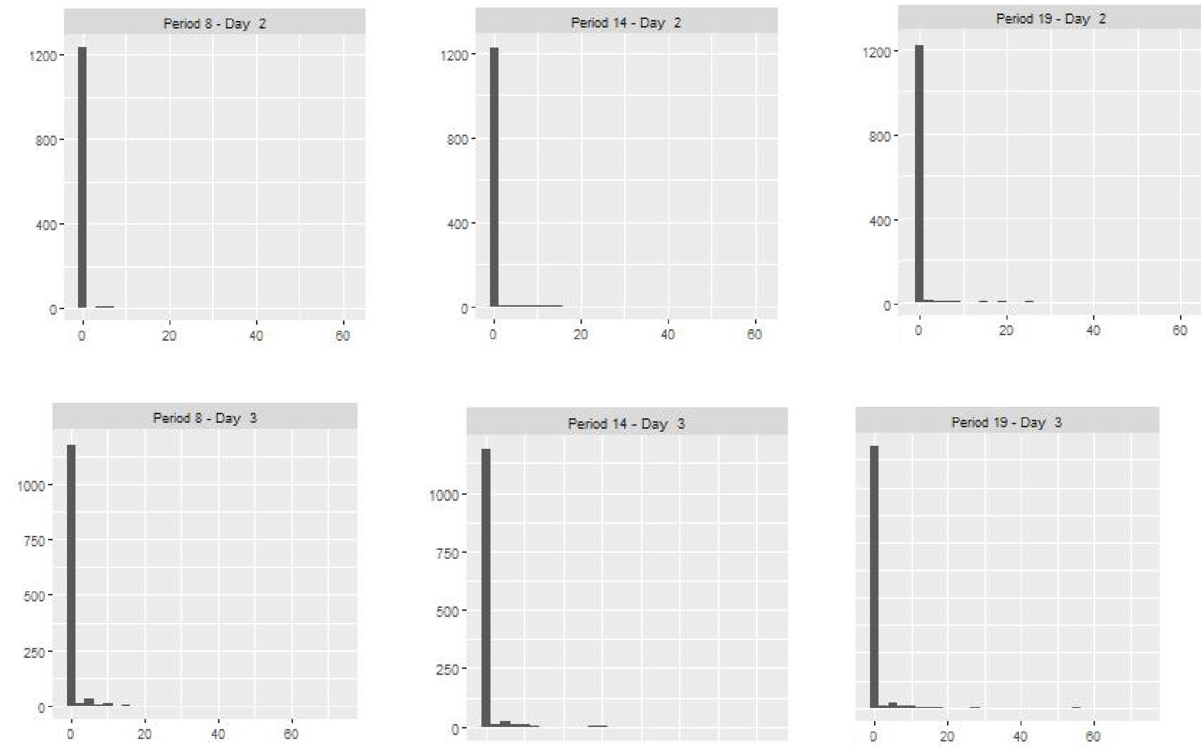


Figure 73. Sunday, 2nd of June 2019 (up) and Mid-day, 3rd of June 2019 (down) histograms of the Number of Trips. Severe asymmetry and considerable skewness can be observed in the histograms for the periods of interest.

The distribution of values is skewed towards the empty cells with zero values. In fact, in some cases the frequency of the value 0 is almost one (or in absolute value almost the total number or zones). This pattern is repeated in all hours of all days. The only difference are the values in the x-axis, which account for up to 60 trips between the same Origin and Destination in one hour for the Mid-day group, while on Sundays the maximum is around 25 trips (Figure 73).

3.1.5.2. Results

Using the above 5 groups of days and the 3 peak periods, a total of 15 pairs of OD matrices are generated, having in each pair a matrix estimated using the mean and a matrix estimated using the ε-med. It can be observed that more cells are populated in the case of the ε-med.

The 3D representation of the two obtained matrices for one hour are presented below.

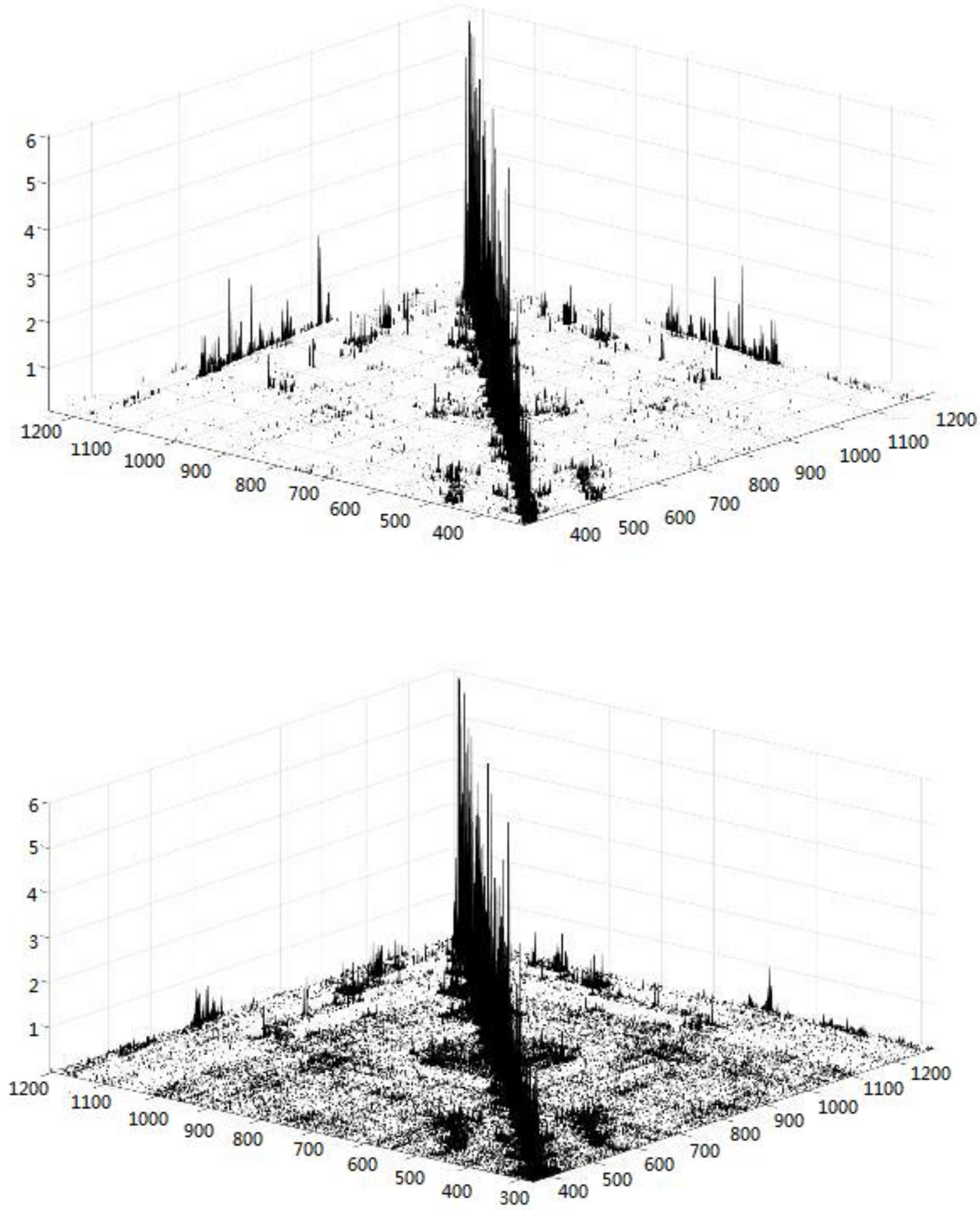
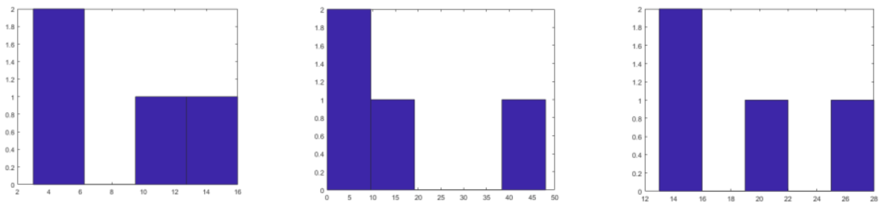


Figure 74. Representative hourly matrix using the mean (up) and the ε-med (down)

Histograms of a random OD pair are presented to demonstrate the skewness as well as the mean value and the e-med value for the same vector. This vector contains the trip numbers for each day of every group that corresponds to a specific OD element in the matrix. It will be proven that the e-med algorithm stirs the local estimate towards the skewness direction.

Mean value: 11.5

e-med value: 10.5



Mean value: 9

e-med value: 9.5

Mean value: 16.5

e-med value: 24

Mean value: 19.25

e-med value: 20.5

Figure 75. Histogram of random OD element trips for the group of Fridays for period 8, 14, 8 and 9.

Below, tables with a set of statistical data for each of the five groups and each of the 3 periods are provided containing the total number of trips, the percentage of zeros, the average value, the average value excluding the zeros and the maximum values.

	8:00-9:00		**14:00-15:00**		**19:00-20:00**	
	mean	**e-med**	**mean**	**e-med**	**mean**	**e-med**
Total number of trips	1,124,900	1,543,390	1,341,326	1,810,575	1,123,000	1,539,800
% of zeros	83.59	83.59	81.90	81.90	84.06	84.06
Average number of trips	0.7097	0.9737	0.8462	1.1423	0.7085	0.9714
Average number of trips without zeros	4.32	5.93	4.67	6.31	4.44	6.09
Maximum number of trips	532.25	532	827.75	757	666.25	662.5
Conditional Entropy	0.86	0.78	0.97	0.88	0.84	0.76
Improvement of Conditional Entropy	9%		9%		9%	

Table 25. Summary of the comparison results for the group Fridays

	8:00-9:00		14:00-15:00		19:00-20:00	
	mean	e-med	mean	e-med	mean	e-med
Total number of trips	1,124,900	1,543,390	1,047,900	1,452,159	1,199,570	1,639,000
% of zeros	83.59	83.59	84.70	84.70	84.05	84.05
Average number of trips	0.7097	0.9737	0.6611	0.9161	0.7568	1.0340
Average number of trips without zeros	4.32	5.98	4.67	6.39	4.74	6.48
Maximum number of trips	532.25	532	643.5	642.5	666	723.5
Conditional Entropy	0.8	0.72	0.87	0.79	0.84	0.76
Improvement of Conditional Entropy	9%		9%		9%	

Table 26. Summary of the comparison results for the group Last Weekdays

	8:00-9:00		14:00-15:00		19:00-20:00	
	mean	e-med	mean	e-med	mean	e-med
Total number of trips	1,208,100	1,110,700	1,292,200	2,732,200	1,267,800	2,690,200
% of zeros	72.37	87.94	71.99	71.99	72.97	72.97
Average number of trips	0.7621	0.7007	0.8153	1.7237	0.7998	1.6972
Average number of trips without zeros	2.75	5.80	2.91	6.15	2.95	6.27
Maximum number of trips	599.33	615.5	814.89	823.5	716.11	705
Conditional Entropy	2.5	2.39	2.55	2.43	2.44	2.33
Improvement of Conditional Entropy	5%		5%		5%	

Table 27. Summary of the comparison results for the group Mid-Weekdays

	8:00-9:00		14:00-15:00		19:00-20:00	
	mean	e-med	mean	e-med	mean	e-med
Total number of trips	501,530	723,819	919,830	1,263,000	927,550	1,274,860
% of zeros	91.19	91.19	87.32	87.32	87.20	87.20
Average number of trips	0.3164	0.4566	0.5803	0.7968	0.5852	0.8043
Average number of trips without zeros	3.59	5.18	4.57	6.28	4.57	6.28
Maximum number of trips	304.25	328	732.75	730.5	1177.5	2327
Conditional Entropy	0.43	0.39	0.66	0.6	0.67	0.61
Improvement of Conditional Entropy	9%		9%		9%	

Table 28. Summary of the comparison results for the group Saturdays

	8:00-9:00		14:00-15:00		19:00-20:00	
	mean	e-med	mean	e-med	mean	e-med
Total number of trips	394,170	747,626	823,930	1,419,185	873,930	1,492,620
% of zeros	90.46	90.46	85.14	85.14	84.65	84.65
Average number of trips	0.24	0.47	0.51	0.89	0.55	0.94
Average number of trips without zeros	2.60	4.94	3.49	6.02	3.59	6.13
Maximum number of trips	296.5	285	911.5	962	699.67	710.5
Conditional Entropy	0.58	0.55	0.98	0.92	1.02	0.96
Improvement of Conditional Entropy	6%		6%		6%	

Table 29. Summary of the comparison results for the group Sundays

As we can see in these tables, the total number of trips mostly increases, arriving in some case to double the number of trips in the e-med matrix in comparison to the mean matrix (this requires further research since the average number of trips per person would increase significantly). The number of zeros is the same since it is related to cells always having zero, so in both cases, the result is zero. With regards to the average, as expected increase as the total does while the situation in the case of the maximum number of trips in the matrix does not present a clear pattern. Finally, with regards to the conditional entropy in every scenario, it is smaller in the case of the e-med. This indicates that there is greater similarity between the e-med values of the OD matrix and the set of real OD matrix values than in the mean case, thus supporting the statement that it is more representative. The improvement varies between 5 and 9% and it seems that it is more or less constant in every group of days, independently of the hour (for the hours analysed).

3.1.6. Validation of the structural similarity measure

The aim of this section is to validate the structural similarity measure of OD matrices proposed in MOMENTUM and described in Section 2.3. To do so, we will use data from the Madrid case study and, more specifically, OD matrices extracted from mobile phone data. The section is structured as follows. First, we will show the validation approach used, then we will describe the results obtained, and finally, we will discuss the main conclusions drawn from this study.

3.1.6.1. Validation approach

The validation has consisted in checking the capability of the proposed similarity measure to quantify previously known seasonal and intra-week differences in mobility patterns and to detect outliers, both due to events modifying standard mobility patterns and to distortions in the input data. A set of hourly OD matrices extracted from mobile network data from a large number of days is used.

3.1.6.1.1. Approaches for day-to-day comparison from the similarity measure

For the validation of the similarity measure, an intra-month comparison has been chosen, that is, for each of the months mentioned above, the mobility patterns of each day have been compared with the rest of the days of the same month. This day-to-day comparison has been carried out using two different approaches for:

- **Approach Pairwise-Daily:** The first approach consists of the pairwise comparison of daily aggregated OD matrices for each month. For example, in October 2018, the daily aggregated OD Matrix for October 1st is compared with the one from October 2nd, 3rd, etc. Furthermore, for each considered month, the structural similarity values are normalized to facilitate its visualization. The normalization applied is the following:

$$NormSim(OD_i, OD_j) = \frac{Sim(OD_i, OD_j) - mean(Sim(OD))}{sd(Sim(OD))}$$

  where mean(Sim(OD)) and sd(Sim(OD)) are the mean and the standard deviation of the similarity of all day-by-day comparisons in the month considered, respectively.

- **Approach Pairwise-Hourly:** In this second approach, instead of comparing mobility between days using a single OD matrix, this is done by comparing in pairs, the 24 hourly OD matrices of the days under consideration. As an example, to compare the mobility of October 1 vs. October 2, the 12am hourly OD matrix of October 1 is compared with the 12am matrix of October 2, the 1am matrix of October 1 is compared with the 1am matrix of October 2, and so on. In this approach, the similarity between two days is given by the weighted average of the similarity between the hourly matrices of both days. More formally, the structural similarity between day i and day j is calculated as follows:

$$Sim_{ij} = \frac{1}{|P|}\sum_{p \in P} w_{ij}^{p} NormSim(OD_i^p, OD_j^p)$$

  Where:

  - P is the number of periods in the day (24)

- ω_{ij}^{p} is the weight that measures the importance of the period p for the comparison of days i and j, being trips $_i^p$ the total number of trips of the hourly OD matrix of day i at period p (analogous for day j)

$$w_{ij}^{p} = \frac{\max\left(trips_i^p, trips_j^p\right)}{\sum_{p \in P} \max\left(trips_i^p, trips_j^p\right)}$$

3.1.6.1.2. Objectives

The objectives of the experimentation carried out are the following:

- Assess the applicability of the measure of structural similarity of OD matrices developed in MOMENTUM. This will be done by checking to what extent the measure is confirming a series of mobility patterns that are usually taken for granted, at least in the context of Madrid Case Study. The following patterns have been assessed with the measure:

  (1) Strong difference in structure and mass between weekdays and weekends.
  (2) Strong similarity in structure and mass from Tuesday to Thursday.
  (3) Some differences in Mondays compared to the rest of weekdays.
  (4) Differences in Fridays compared to the rest of weekdays, especially in the evening: more trips from 8pm to the end of the day (due to leisure activities); instead of three home-work peak hours (morning, midday and evening), only two (morning and midday)
  (5) The same days of the week (Monday, Tuesday, etc.) may show some similarities across weeks, due to recurrent activities.
  (6) Differences between weekdays and weekends and each of them inner similarities are expected to be stable all months.
  (7) 'Season transition' months such as June or September present differences throughout the month.
  (8) Pre-bank holidays days show similar patterns to Fridays.

- Analyse both approaches to compare day-to-day mobility patterns (pairwise-daily vs pairwise-hourly) to study which of the two approaches is more appropriate for observing the patterns mentioned above
- Analyse whether the structural similarity measure allows for the identification of other patterns not mentioned above such as events or incidents in mobility.

According to these objectives, the days of the weeks will be compared first to each other, without differentiating between months, to verify if the patterns (1) - (5) are met. Then a similar analysis will be conducted but separated by months, to verify the robustness of the structural similarity measure to identify the patterns (1) - (6) in months with different characteristics. Finally, a day by day comparison for some of the months considered will be examined, with the main objective of verifying the patterns (7) and (8) are observed.

3.1.6.1.3. Input origin-destination matrices

To validate the structural similarity measure of OD matrices, an official aggregated version of the most detailed study zoning in Madrid region has been used. This is composed of 208 zones[1] which can be seen in Figure 76 ("208 zoning"). In this way, the compared matrices have a dimension of 208x208. As for the periods, we used the hourly

[1] https://datos.crtm.es/datasets/zonificacionzt208/data?geometry=-9.054%2C39.790%2C1.400%2C41.252

OD matrices from all the available months of study: October 2018, February 2019, April 2019, June 2019, July 2019, October 2019, February 2020. Thus, a total of (31+28+30+30+31+31+29) x 24 = 5,040 OD matrices have been used for the analysis.

Figure 76. The 208 zoning established by the regional transport authority (CRTM) in Madrid region.

3.1.6.1.4. Implementation

The experimentation has been carried out using the Colab[2] in its version Pro. Colab is a platform developed by Google for programming and running Python scripts, and its main feature is that it offers GPU access for free or at a very low cost. The GPU assigned by Google Colab for the experimentation has been the NVidia Tesla V100-SXM2-16GB[3] with compute capability 7.0. The implementation has been done using version 2.0 of the TensorFlow library[4].

3.1.6.2. Results

3.1.6.2.1. Global comparison among days of the week

As mentioned above, in this first analysis the aim is to validate whether the OD matrix structural similarity measure verifies patterns (1) - (5) and to what extent each of the two approaches to compare day-to-day travel patterns provides more actionable results.

To this end, the distribution of the structural similarity measure has been analysed for each day of the week against the rest based on the intra-month comparison. That is to say, for the month of October 2018, 31x31=961 comparisons have been made which result from the comparison of all vs all of the 31 days of this month. This procedure was extended to the rest of the analysed months in a similar way to obtain 6,308 measures of structural similarity, both for the pairwise-daily approach and for the pairwise-hourly approach. These measures have been grouped according to the days of the week of the two dates compared. Figure 77, Figure 78, Figure 79 and Figure 80 show the distribution of the structural similarity of the mobility patterns of each day of the week against the

[2] https://colab.research.google.com/
[3] https://www.nvidia.com/es-es/data-center/tesla-v100/
[4] https://www.tensorflow.org/guide/effective_tf2

rest using box diagrams, for the pairwise-daily and pairwise-hourly approaches, respectively. More specifically, each panel corresponds to one day of the week and each box to the distribution of similarity both with respect to the same day of the week and to the rest. In order to interpret the results, it is convenient to know that in the proposed measure the lower the value the higher the similarity between the compared OD matrices.

In order to avoid biases in the comparison, the following days associated with bank holidays have been removed for the analysis done in this section:

- October 2018: 11, 12, 13, 14 and 31
- April 2019: 12, 13, 14, 15, 16, 17, 18, 19, 20, 21 and 22
- October 2019: 11, 12, 13 and 31

Figure 77 and Figure 78 show interesting results related to the similarity between the days of the weekend. As can be seen, both Saturdays and Sundays are more similar to each other than to the rest of the days of the week. This would partially validate the measure for the pattern (5), at least for weekend days. Furthermore, the difference between these two days and the weekdays is much larger than the one observed between the weekdays themselves, thus confirming the pattern (1). Another interesting aspect is the fact that the similarity between a Saturday and a Sunday (and vice versa) is somewhere in between the Saturday (and Sunday) and the weekdays, a fact that is to be expected given that Saturdays have a larger proportion of people working than Sundays. This is true for both comparison approaches. If weekdays are examined, it can be seen that they are all very different to Saturdays and Sundays, although to a lesser degree to Saturdays than from Sundays. This pattern is also to be expected and contributes to the validation of the measure.

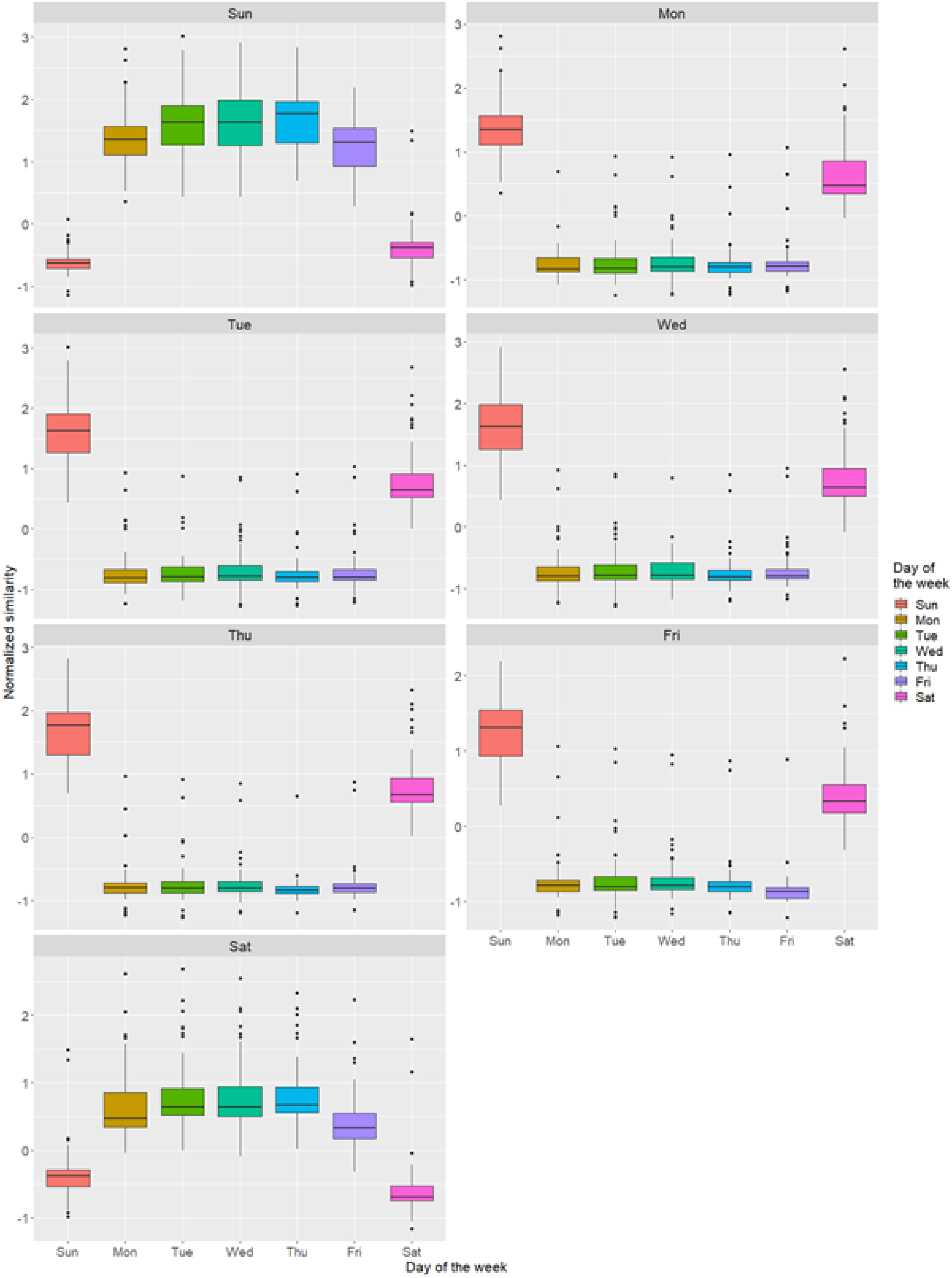


Figure 77. Distribution of the similarity for all days of the week (Approach pairwise-daily)

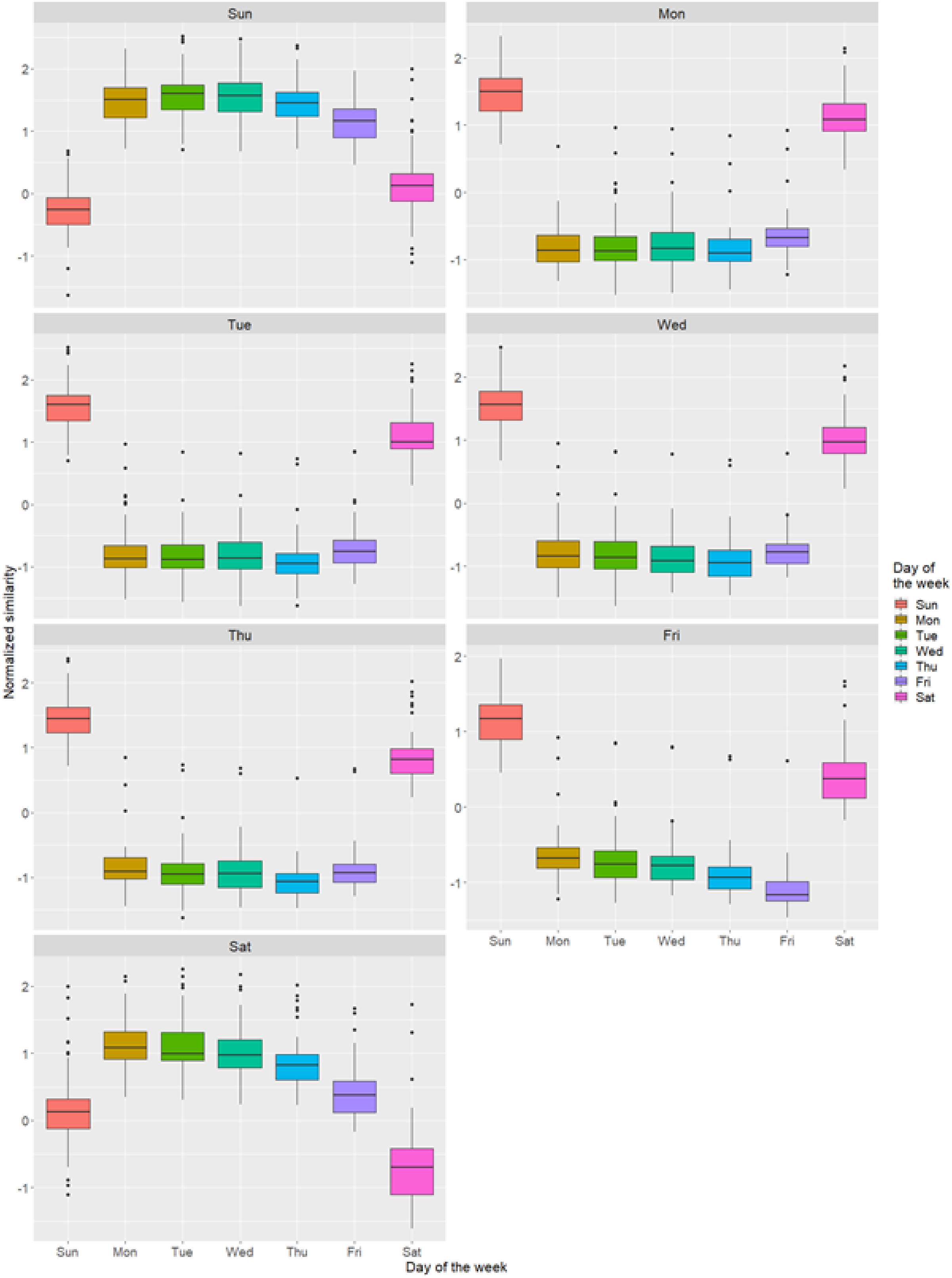


Figure 78. Distribution of the similarity for all days of the week (Approach pairwise-hourly)

To enable a more focused analysis of the patterns observed for the weekdays, Figure 79 and Figure 80 show the same type of graphs, but only for the weekdays. The first aspect to highlight is that the pattern (2) can be clearly seen with both comparison approaches (daily and hourly), as there is a high similarity among Tuesdays, Wednesdays and Thursdays.

As far as the pattern (3) is concerned, it is not seen in the box diagram related to Monday, with none of the two approaches. However, looking at the panels for Thursday and Friday (as well as those for Saturday and Sunday in Figure 77 and Figure 78), it is clear that Mondays are different from the rest of the weekdays when compared to these days. This is most clearly seen with the pairwise-hourly approach. This would indicate that there are some indications that the measure is meeting the pattern (3), although further analysis would be needed to confirm this statement as explained above. It may be the case that this pattern is not as strong as assumed.

As for the pattern (4), this can be clearly seen with the pairwise-hourly approach. However, with the pairwise-daily approach, the differences on Friday with the other days are not so clear. This may be due to the fact that trips on an aggregate level are not so different between Fridays and the rest of the weekdays. However, their temporal variation is remarkable. As mentioned above, Fridays have only two peak periods instead of three (but with a more pronounced afternoon peak) and a greater number of trips after 8pm, mainly related to leisure. This shows that to observe certain types of patterns which imply a temporal variation of demand that may not be observable on an aggregated level, the pairwise-hourly approach is more appropriate than the pairwise-daily approach. Analysing the pattern (5), it can be seen that the pairwise-hourly approach allows us to clearly see it in the Thursday and Friday panels (Figure 80), as the similarity values of these two days with respect to each other are lower than with respect to the other days. For the rest of the days of the week, this trend is not observed. However, doing the same exercise as for the pattern (3), it is possible to see that when Mondays, Tuesdays and Wednesdays are compared with Thursdays, Fridays, Saturdays and Sundays differences are observed between them, which somehow gives indications that this pattern may exist, but this is unobservable using only the similarity measure. As it happened with pattern (4), the pairwise-daily approach does not show a clear pattern (5). This confirms the idea that a more aggregated approach may not be appropriate to identify temporal differences in travel patterns.

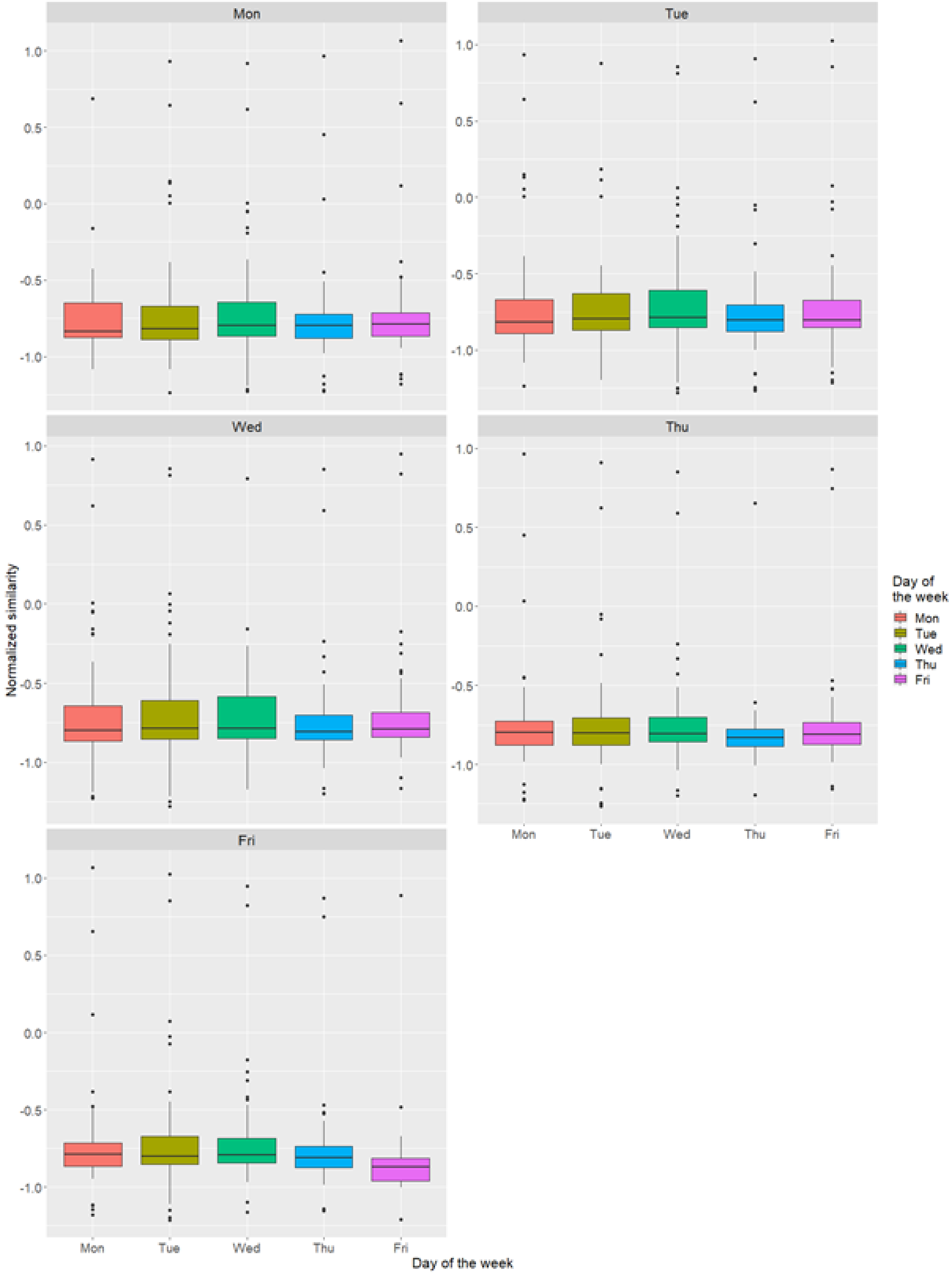


Figure 79. Distribution of the similarity for weekdays (Approach pairwise-daily)

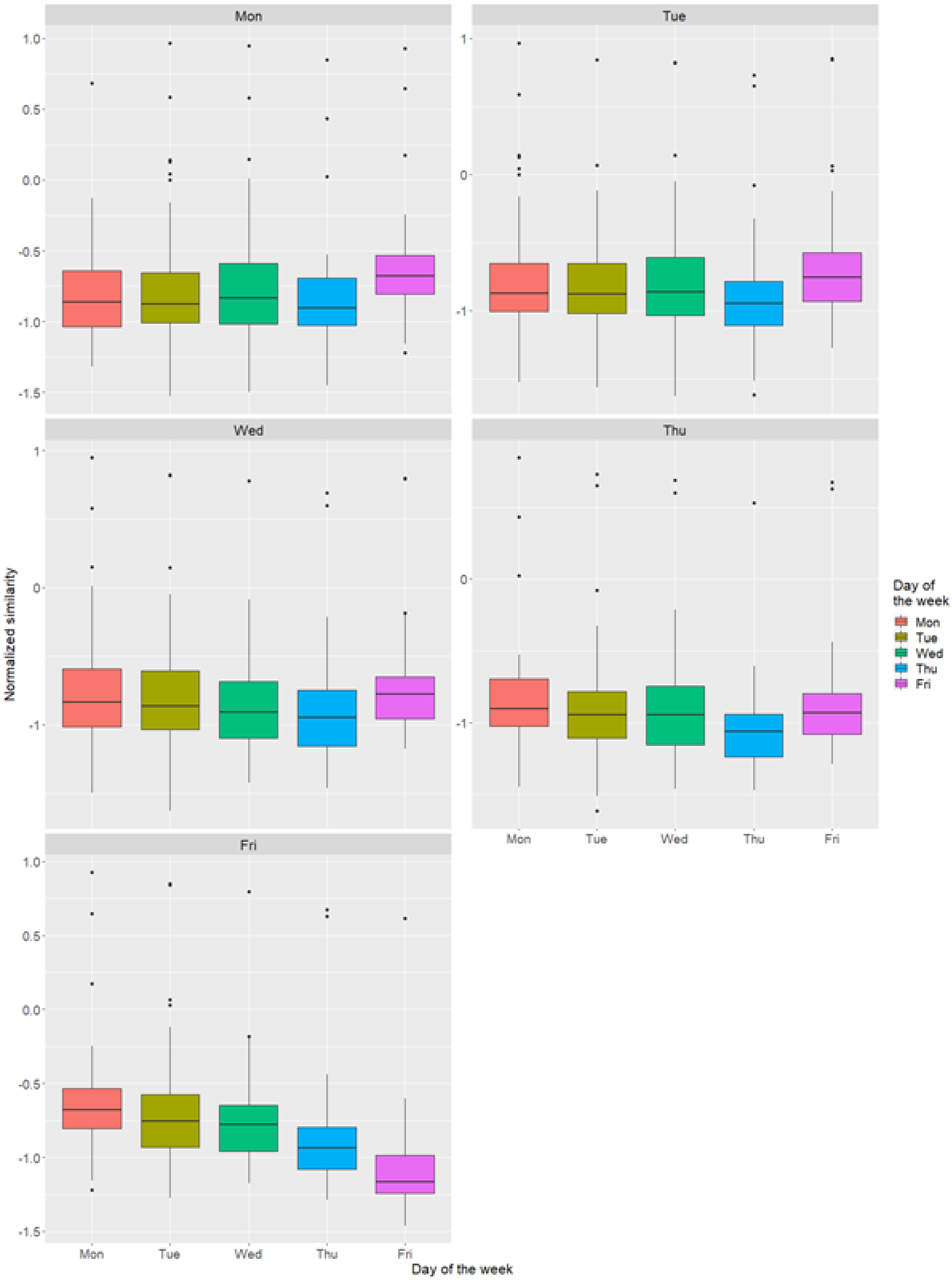


Figure 80. Distribution of the similarity for weekdays (Approach pairwise-hourly)

Figure 81. Monthly distribution of the similarity for all days of the week (Approach pairwise-daily)

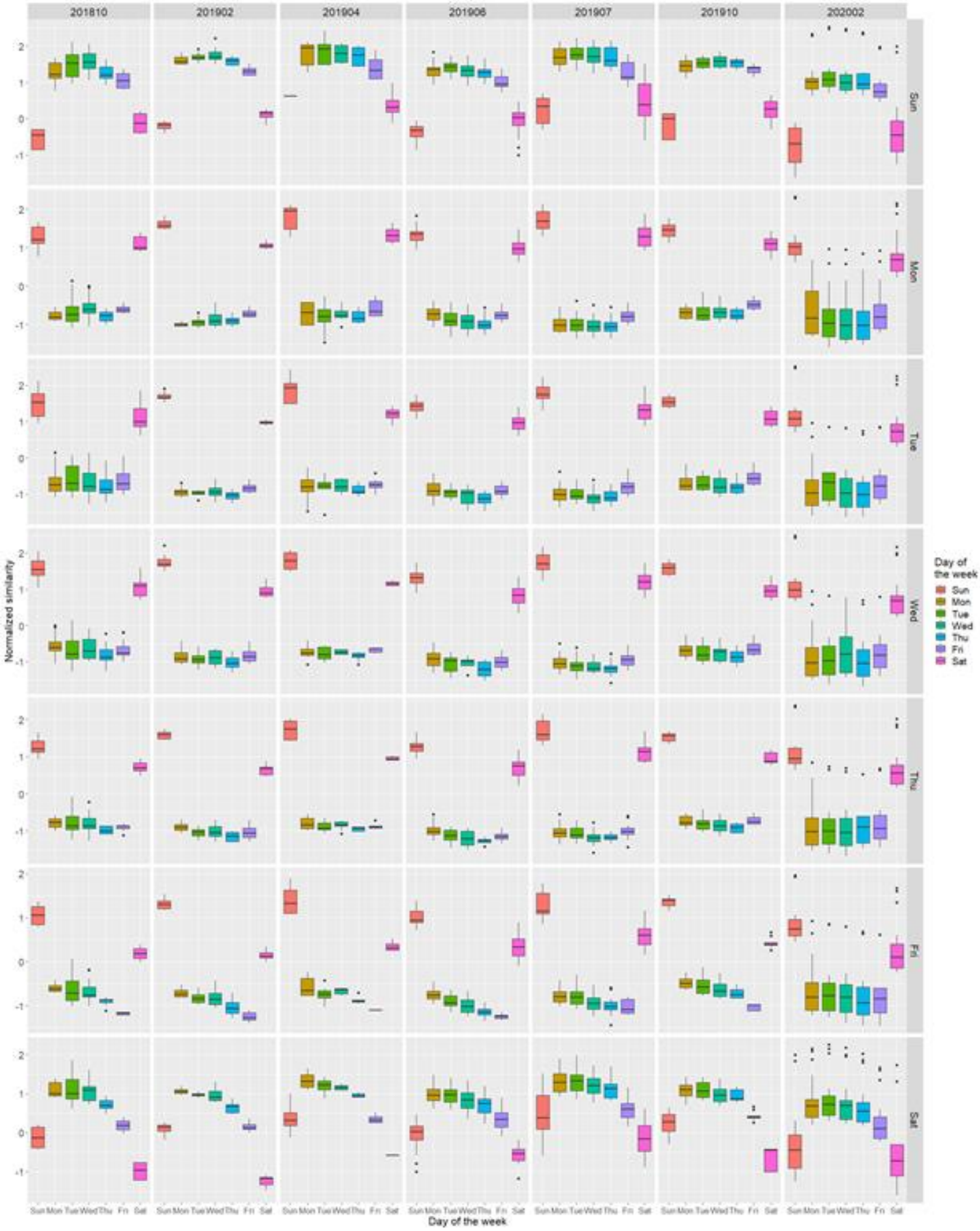


Figure 82. Monthly distribution of the similarity for all days of the week (Approach pairwise-hourly)

3.1.6.2.2. Monthly comparison among days of the week

This second block of results shows a similar analysis to the one presented in the previous subsection, but separating it by months, to see to what extent the patterns (1)-(5) are also observed in each month, as well as validating whether the measure complies with the pattern (6). The results are displayed in Figure 81, Figure 82, Figure 83 and Figure 84.

Firstly, it can be seen that both pattern (1) and (2) are appreciated with both comparison approaches, except for pattern (2) in the month of October 2018. The reason for this exception may be due to the particular characteristics of this month and to possible errors in the generation of the October 9th and 10th matrices, which could have distorted the results. We will return to this point in the next section.

As for the pattern (3), it can be observed in October 2018 and February 2019 with both comparison approaches, but not for the rest of the months. This may be due to the fact that the other months present elements such as holidays, vacation period, etc. which may alter this relationship. This is further explored in the day-to-day comparison reflected in the next section.

Continuing with the pattern (4), the charts show it clearly in all months, especially when using the pairwise-hourly approach, for the same reasons as stated above, when comparing all the periods at once. Only for the month of February 2020, it is not appreciated with the pairwise-daily approach. The probable reason for this is the high variability of the similarity between days that we observe in this month. A more in-depth analysis (similar to the ones that we will show in the next section) showed that the mobility patterns of the last two weeks of this month are very different for the one of the first two weeks, considering both weekends and weekdays. In our opinion, the variation in mobility patterns in the last two weeks of February may be due to the irruption of the Covid-19 in Madrid this month.

With regard to the pattern (5), the results are analogous to those presented in the previous section, in the sense that it is partially fulfilled, i.e., not for all days of the week. The pairwise-daily approach allows us to observe this similarity between the same days for Fridays, Saturdays and Sundays in a stable way in all months except for Fridays in February 2020. The pairwise-hourly approach shows this pattern for Thursdays, Fridays, Saturdays and Sundays more clearly and also very stably for all months except for Thursdays in February 2020.

Finally, pattern (6) can be clearly observed in the results from all months, regardless of the approach used.

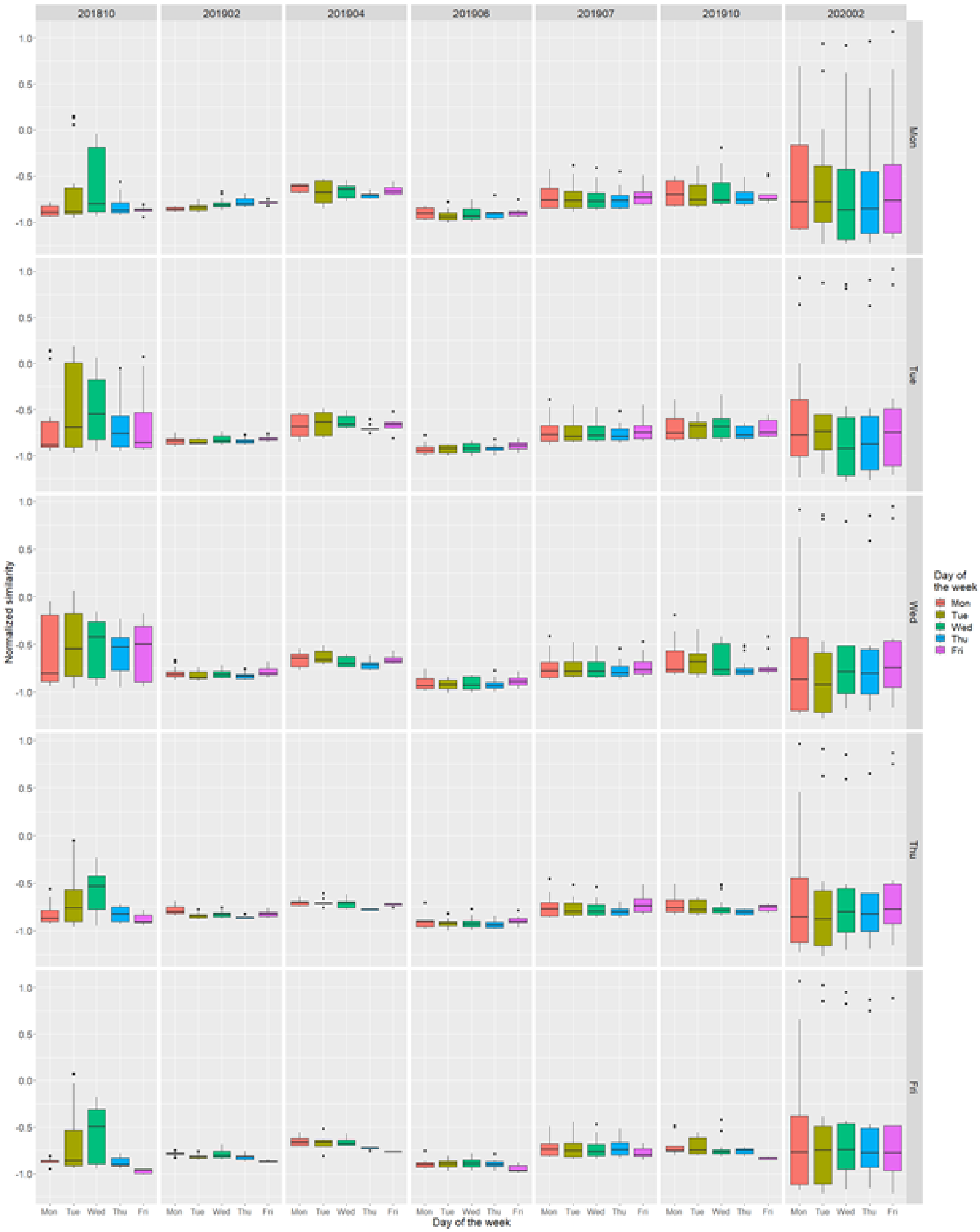


Figure 83. Monthly distribution of the similarity weekdays (Approach pairwise-daily)

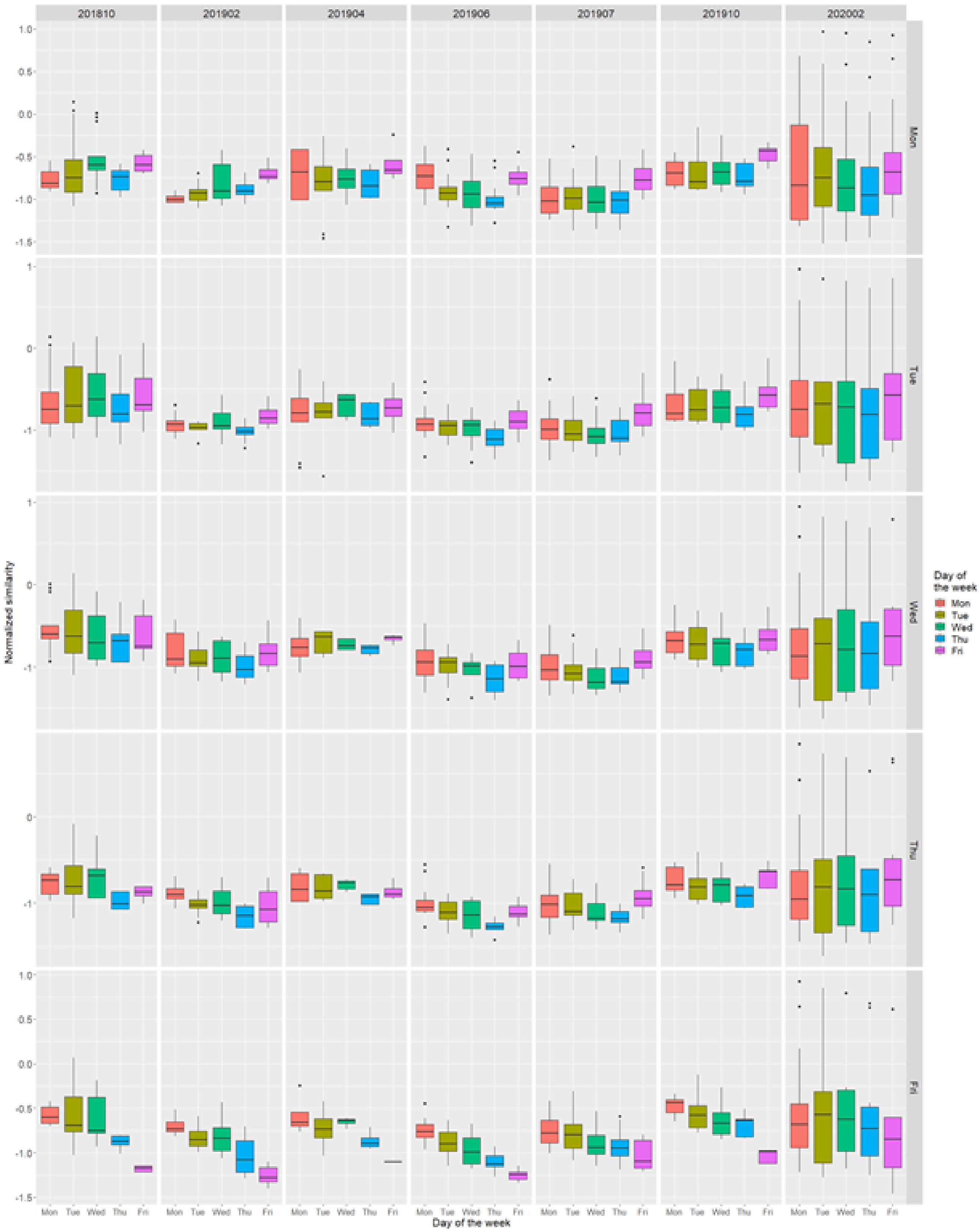
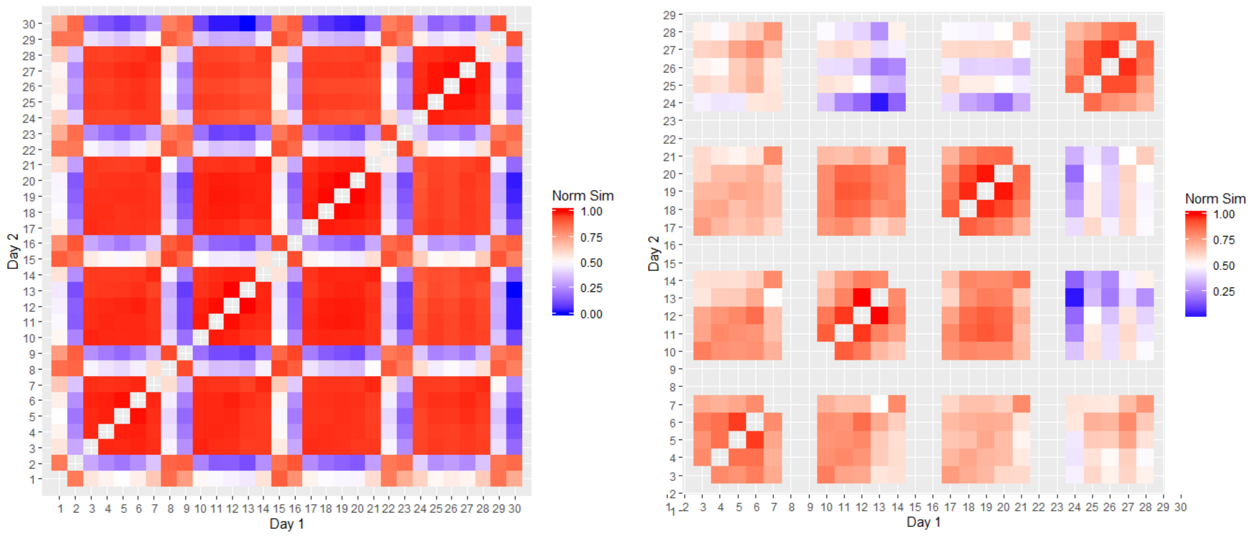


Figure 84. Monthly distribution of the similarity weekdays (Approach pairwise-hourly)

3.1.6.2.3. Day to day comparison

The aim of this section is to use day-to-day comparisons with two purposes: (1) to analyse whether the similarity measure meets the patterns (7) and (8) indicated at the beginning of the section, and (2) to assess if the similarity measure is able to show other relevant patterns or to detect anomalies in travel patterns. The first objective is done by examining the day-by-day comparison, for the months of October 2018 and 2019, and June 2019. The results of this comparison for the pairwise-daily approach are shown in the form of heat diagrams in Figure 85, Figure 87 and Figure 89 respectively, comparing all days of the week with each other (figures (a)), and comparing only weekdays to improve the visualisation of certain patterns (figures (b)). Figure 86, Figure 88 and Figure 90 are analogous but for the pairwise-hourly approach. In these heat diagrams, the colour scale is given by the similarity between the days being compared. Dark blue colour indicates a lower similarity, white colour an intermediate similarity and dark red colour a high similarity. As the similarity values are standardised month by month, it is not appropriate to compare the colour scale from one month to another, nor between figures (a) and (b), as they may represent different similarity values.

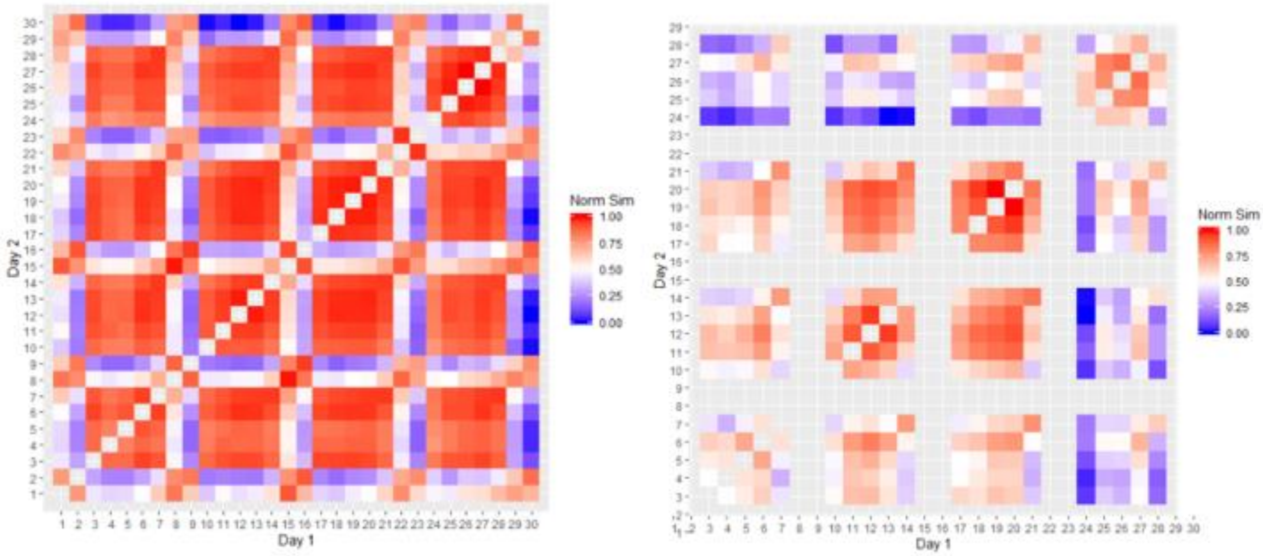


(a) All days

(b) Only weekdays

Figure 85. June 2019 with approach pairwise-daily

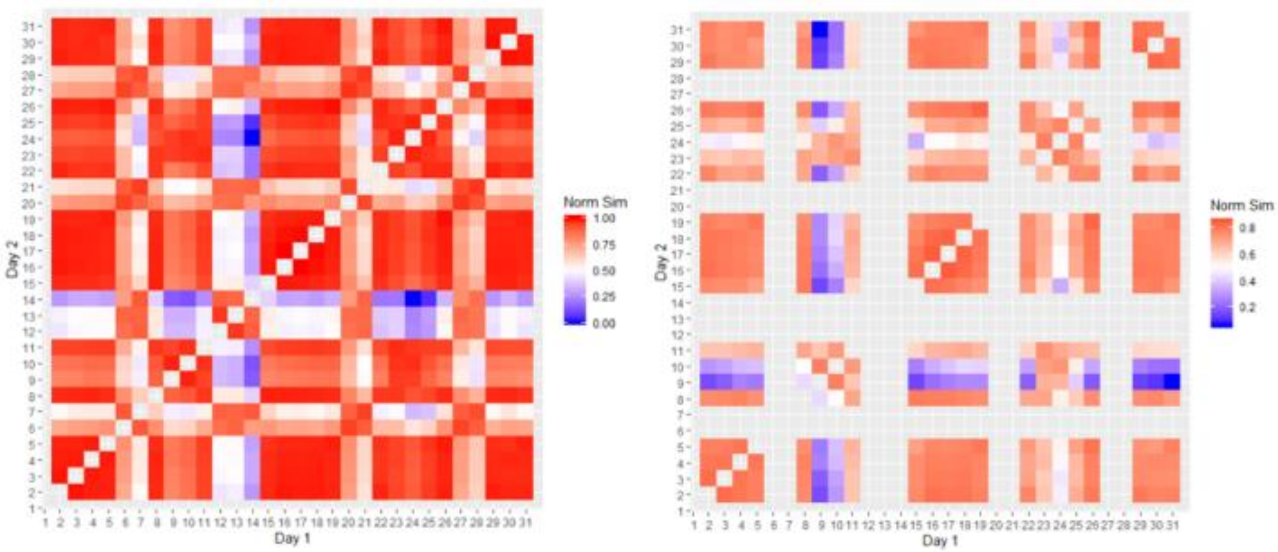
Figure 85 and Figure 86 can be used to validate the pattern (7). They correspond to the day-to-day comparison for the month of June 2019 using the daily and hourly approaches, respectively. Here it can be seen that the pattern of "transition month" is very clear in figures (b) where only weekdays are compared. It can be also seen how the first two weeks are different from the last two weeks, as there are usually more people who have gone on holiday and also the school period has ended. In figures (a), with the pairwise-daily approach, this transition cannot be seen at the weekends, whereas with the hourly approach it can be noticed. If the first two weeks are analysed, it is possible to see that the strip in which weekdays are compared to weekends becomes darker as we move to the right or upwards, i.e., as dates move forward in the month, indicating that mobility patterns at weekends are also different. Thus, it is possible to conclude that the proposed similarity measure also quantifies and characterises the pattern (7), especially when the pairwise-hourly comparison approach is applied.



(a) All days

(b) Only weekdays

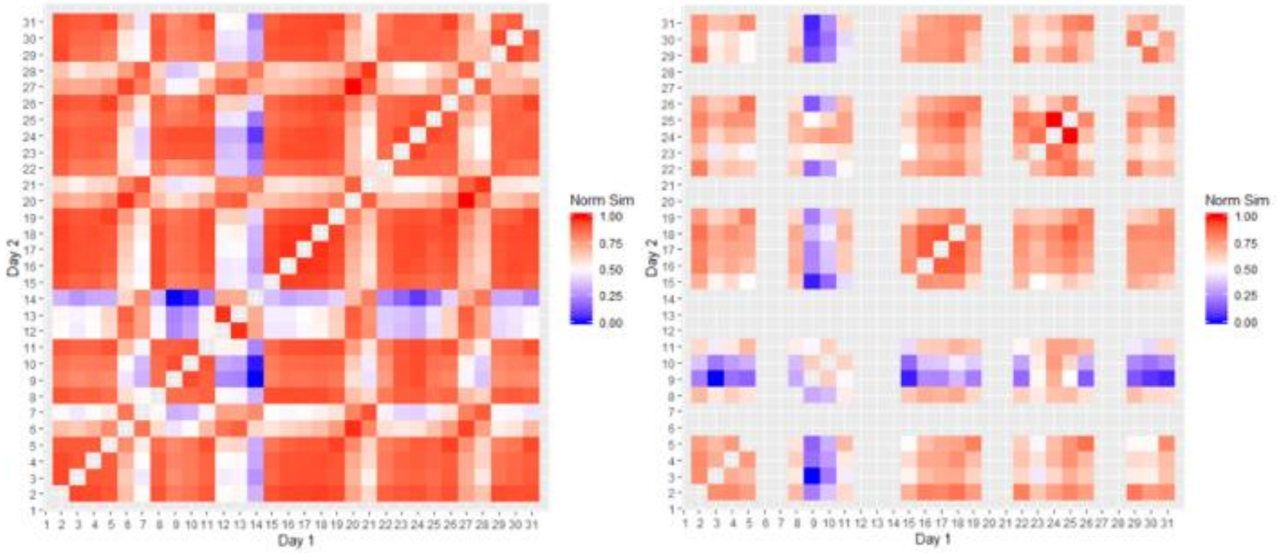
Figure 86. June 2019 with approach pairwise-hourly



(a) All days

(b) Only weekdays

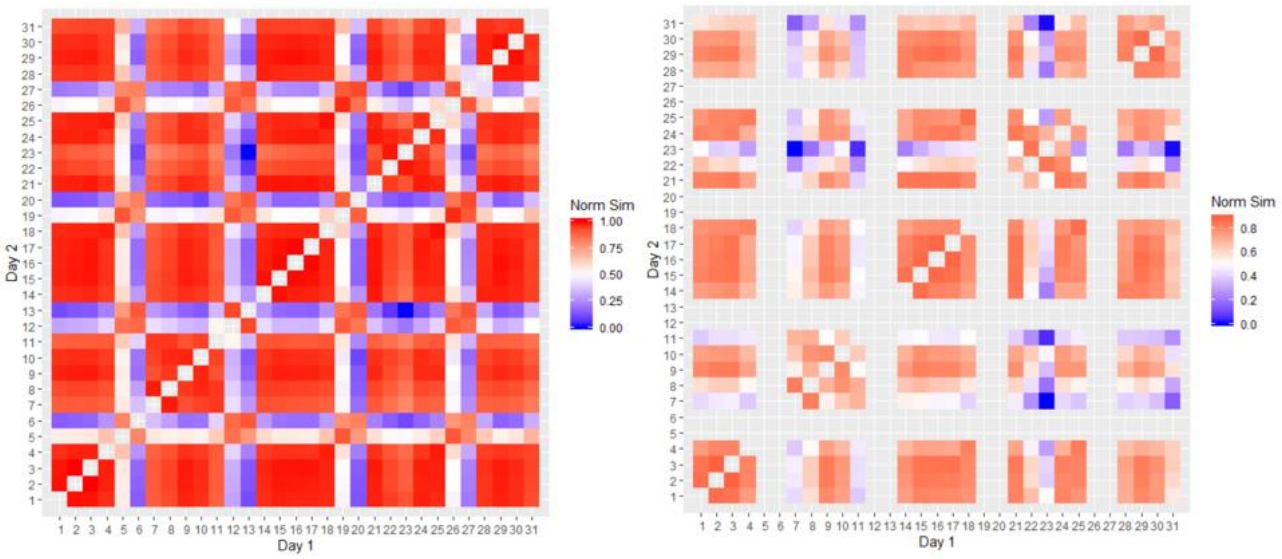
Figure 87. October 2018 with approach pairwise-daily



(a) All days *(b) Only weekdays*

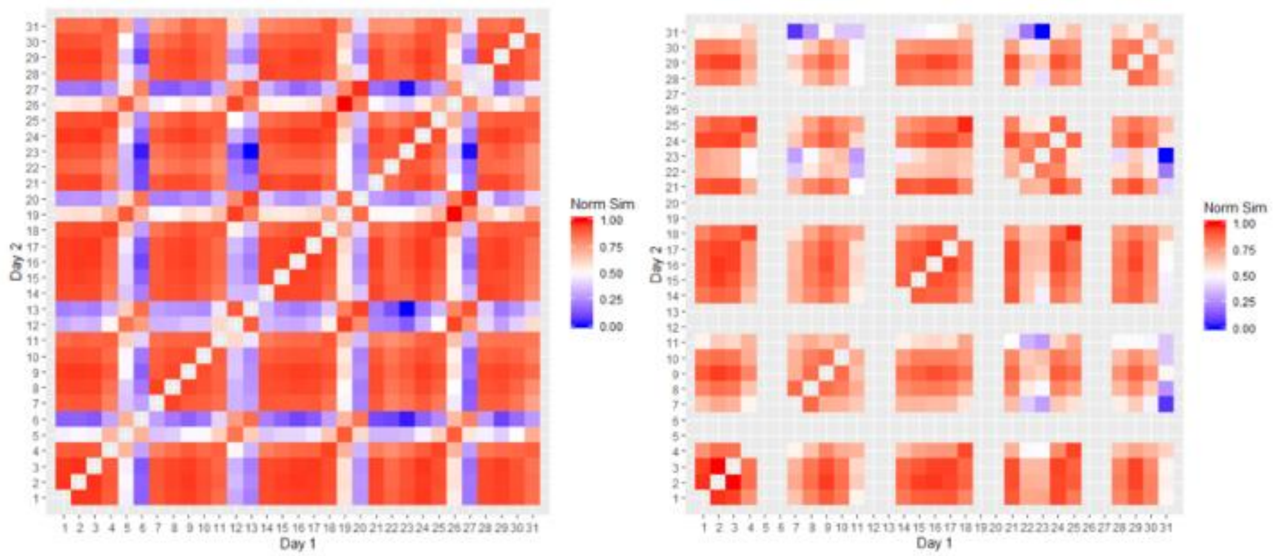
Figure 88. October 2018 with approach pairwise-hourly

To conclude validation of the structural similarity measure, the analysis focuses on the last pattern, number (8), which refers to "Pre-bank holidays days show similar patterns to Fridays". To do this, the OD matrices of October 2018 and October 2019 will be analysed (Figure 87-Figure 88 and Figure 89-Figure 90, respectively), as these months have two pre-bank holidays, specifically 11th and 31st October, since October 12th and November 1st are national holidays in Spain. Starting with October 2018, it is possible to see in Figure 87 and Figure 88 that both approaches capture the bank holiday of October 12th, and show that the mobility patterns of these three days are very different from the rest. To observe the pattern (8), it is useful to focus first on the figures (b). Here it can be seen that with both approaches October 11th is different from the rest of Thursdays of the month, and also that it has a high similarity with Fridays (especially with the pairwise-hourly approach). However, we should remark that there is no greater similarity with Fridays than with the rest of the days of the week. If we look at the figures (a), it can be seen that October 11th, despite being a Thursday, shows a similarity with weekends like the one of Fridays, especially when the pairwise-hourly approach is used. Looking at Wednesday, October 31st 2018, in figures (b), a higher similarity can be observed between this day and the Fridays of the month (although the hourly approach shows this pattern more clearly). In October 2019, October 11th is not valid to assess pattern (8) since it is a Friday. However, the pattern (8) can be clearly noticed for Thursday 31st October, especially with the pairwise-hourly approach.



(a) All days *(b) Only weekdays*

Figure 89. October 2019 with approach pairwise-daily



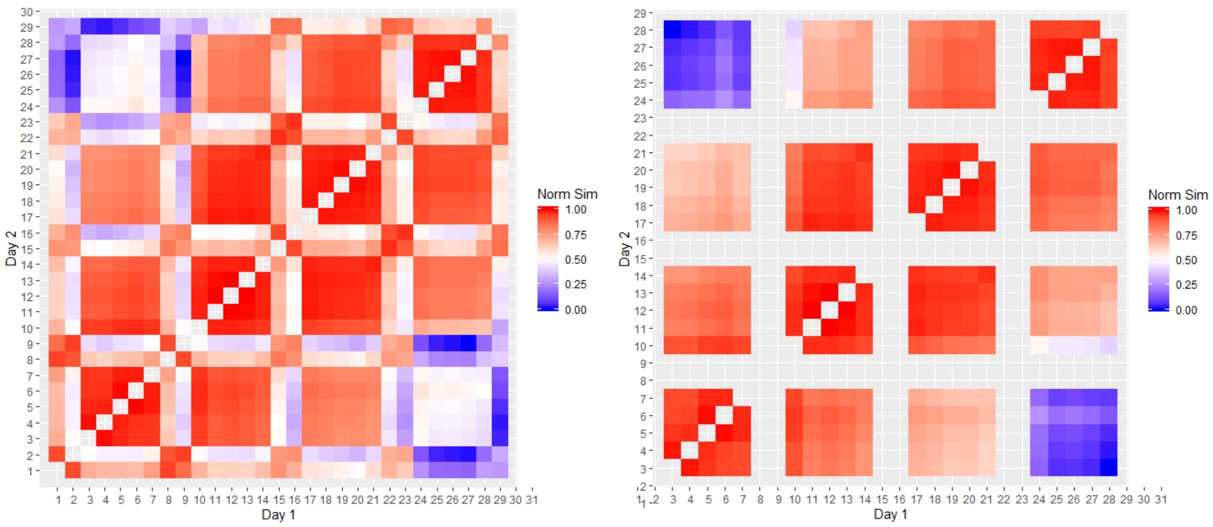
(a) All days *(b) Only weekdays*

Figure 90. October 2019 with approach pairwise-hourly

In this way, we can conclude that the structural similarity measure developed reveals patterns (7) and (8). If we contrast both approaches to comparison, as described above, the pairwise-hourly approach allows both patterns to be observed more clearly, again confirming that it is more appropriate to use it to extract more subtle patterns, as is the case with (7) and (8).

Regarding the second objective of this subsection, we will use the latest graphs again for the months of October 2018 and 2019 to show other interesting patterns revealed by the similarity measure. The first one is the similarity between holidays and weekends, as well as the dissimilarity between these and working days. If we look at the 12th of October 2018, which as we have commented above is a national holiday in Spain, we see that with both approaches, the proposed measure shows how this day is similar to Fridays, Saturdays and Sundays, and at the same time, different to the rest of the days of the week. In the case of 12^{th} October 2019, the pattern observed is also similar. In this case, this bank holiday falls on a Saturday, and what we observe is that this Saturday is more different to the weekdays than the rest of the Saturdays of the month, which confirms that the measure is able to detect that it is not a normal Saturday.

It is also worth highlighting the anomalies that can be observed in these two months. For example, on 23^{rd} and 24^{th} October 2018, and 22^{nd} and 23^{rd} October 2019, both Tuesday and Wednesday are different from the rest of the weekdays. This dissimilarity may be due to the fact that on those days there are UEFA Champions League matches, with Real Madrid playing as home team on 23^{rd} October 2018 and Atletico Madrid on 23^{rd} October 2019. Another anomaly is also observed on October 7^{th} 2019 (more visible with the pairwise-daily approach). The possible reason for this anomaly could be the demonstration that took place that day in Madrid to protest against the inaction of governments on climate change[5], which resulted in traffic cuts. Finally, irregular patterns are also observed on 9^{th} and 10^{th} October 2018. In these cases, no event has been identified that could lead to these irregularities, so these may also be due to some error in the generation of the OD matrices caused by incorrect input data. This shows another possible application of this measure of structural similarity of OD matrices, as is the detection of anomalies in their generation, due for example, to errors in the input data.



(a) All days

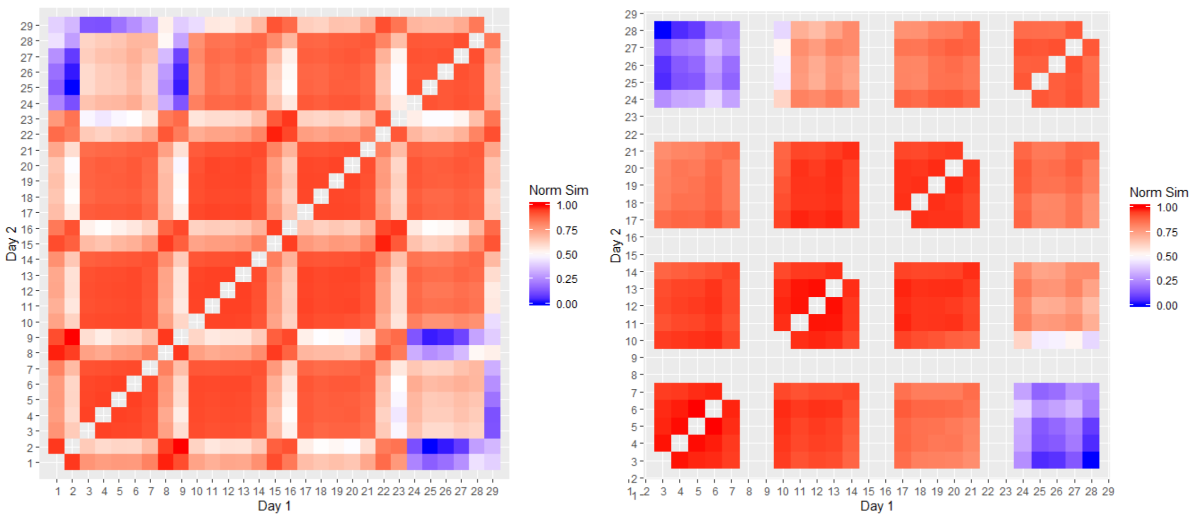
(b) Only weekdays

Figure 91. February 2020 with approach pairwise-daily

To finish with the subsection, we will also show the day-to-day comparison of February 2020 in Figure 91 and Figure 92 for pairwise-daily and pairwise-hourly approach, respectively. With both approaches, we can see that there is a significant anomaly in the last week, although a progressive dissimilarity is also observed in the third week of the month, especially when compared to the first week. In our opinion, this anomaly is due to the changes in mobility patterns in the region of Madrid caused by the first incidents and news of the COVID19 coming from

[5] https://elpais.com/sociedad/2019/10/07/actualidad/1570430675_051785.html

Italy[6] and the discovery of the first cases in Spain[7]. These anomalies can be observed both with the pairwise-daily approach and with the pairwise-hourly approach. Although we have not been able to contrast this trend with other data sources, these results confirm the applicability of the measure of structural similarity developed in MOMENTUM to detect temporary changes in mobility patterns.



(a) All days *(b) Only weekdays*

Figure 92. February 2020 with approach pairwise-hourly

3.1.6.3. Discussion and conclusions

The proliferation of longitudinal data sources describing mobility patterns brings about a demand for procedures capable of detecting deviations from standard values. While outlier detection techniques can be easily translated from other fields to unidimensional measures (e.g., traffic count values), the monitorisation of changes in OD matrices is far from trivial, given the mass-structure trade-off described in Section 2.3 and the spatial information included in them. This has motivated the development of tailored similarity measures in the literature and the improvement of the existing approaches explored in this project.

In this experiment, the OD matrices collected from mobile network data in the context of the Madrid case study served to validate the similarity measure developed in MOMENTUM. Most of the well-known weekly and seasonal mobility patterns have been clearly observed with the results of the measure. The similarity measure was sensitive to patterns such as the differences between weekdays and weekends, the singular behaviour of Fridays and days previous to bank holidays or the differences along weeks in months such as June. This probes the validity of the measure. It seems that pairwise-hourly approach is more descriptive when answering these questions. More concretely, we have seen that the pairwise-hourly approach is more appropriate for comparing daily mobility patterns that in aggregate terms may be similar, but whose temporal variation intra-day is different (e.g. Friday

[6] https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/2/joint-who-and-ecdc-mission-in-italy-to-support-covid-19-control-and-prevention-efforts
[7] https://www.efesalud.com/cataluna-primer-caso-coronavirus-peninsula/

versus the rest of the weekdays or pre-bank holidays). We have also seen that the structural similarity measure can potentially detect anomalies in the mobility patterns due to certain events (e.g. football matches, COVID19) or errors in the generation of the OD matrices.

Apart from this validation, the experiment has also suggested several applications of the similarity measure that are interesting for transport practitioners:

- Quick characterisation of mobility patterns through long periods enables to select a set of dates that may be valuable for obtaining representative OD matrices, complementing the methodology proposed in Section 2.1.
- This can be useful not only for selecting subsets of the analysed data source but also for selecting representative days to conduct fieldwork such as surveys or counts.
- The similarity measure can also detect outlier dates that need further analysis before validating the obtained OD matrix. For instance, as discussed above, Figure 89 and Figure 90 show strong dissimilarities among 22nd October 2019 and 23rd October 2019 data compared to the rest of the month. In this case, this might be the result of anomalies in the input mobile network data or special events in the study area changing mobility patterns.

3.2. Leuven Case Study

3.2.1. Data sources

In the Leuven Case Study, the 2017 City Monitor survey (Category TD_006 in MOMENTUM Data Repository) was used to characterise car-sharing demand patterns. This survey is carried out every three years in different cities of the Flanders region in Belgium, including Leuven. It covers residents over 16 years old, but it only considers people registered in the city. Concretely, 2,669 residents of Leuven, divided into eight different districts, were interviewed between May 2nd and June 6th 2017 in order to collect different indicators on their perception of various issues related to the liveability of the city, among which we can find mobility, along with socio-demographic information.

A more detailed description of the data source can be found in deliverable D3.1. Also, the original survey information is available at https://www.gemeente-en-stadsmonitor.vlaanderen.be/survey

3.2.2. Characterization of car-sharing subscribers

3.2.2.1. Applied methodologies

For the characterization of car-sharing service subscribers, we have focused this analysis on the next question included in the 2017 City Monitor survey: "How many cars, motorcycles/mopeds, bicycles and public transport season tickets does the family you're part of have?" (Question 21 in the original survey). The possible categories of answers were:

- Cars
- Motorcycles, mopeds
- Cycling
- Electric bicycles
- Cargo bikes, bike trailer
- Subscriptions public transport (NMBS, De Lijn)
- Membership car-sharing

Specifically, we will compare those surveyed individuals whose households have at least one car-sharing subscription with those that do not (considered as the control group). These two groups account for 3,704 and 78,180 individuals, respectively. Analysing these two groups according to different socio-demographic, economic and mobility habit variables, the aim is to profile the characteristics that best represent car-sharing service subscribers. However, we have to be aware that there is no information about whether the individual is the subscribers or not, but if there is at least one subscription in the household of the individual. In this way, we are using people living in this type of households as proxies, which do not have to coincide with the subscribers themselves. The variables analysed are sex, age, nationality, labour status, education, incomes, household size and type, the main mode of transport, place of residence, satisfiability with public transport and vehicle ownership.

The number of interviewed individuals is 2,669, which using the expansion factor amounts to a total of 81,884 people.

3.2.2.2. Result analysis

3.2.2.2.1. Sex

The data collected by the survey reflects that there is a higher proportion of female among respondents living in households with at least one car-sharing subscription. Specifically, 57% of those respondents are women. As discussed above, this does not mean that there is a higher proportion of female car-sharing subscribers. In fact, this would go against results that can be found in recent literature, since studies conducted in cities such as Basel in Switzerland (Becker, Ciari, & Axhausen, 2017) or Beijing in China (Yoon, Cherry, & Jones, 2017) showed a higher proportion of male members. However, as we will see in Section 3.2.3, the percentage of women that claims to be already doing car-sharing is higher than that of men, suggesting that there is an interesting particularity in this city that makes car-sharing more attractive for females.

3.2.2.2.2. Age

As can be seen in the following figure, respondents living in households with at least one car-sharing subscription are significantly younger than the control group, indicating that such households are also likely to show a lower average age than the control group households. These results are compatible with those reported in several studies, such as those conducted in the area of San Francisco (United States) (R. Clewlow, 2016) or Basel (Switzerland) (Becker, Ciari, & Axhausen, 2017), indicating that, currently, the attractiveness of car-sharing decreases significantly as the age of the individual increases.

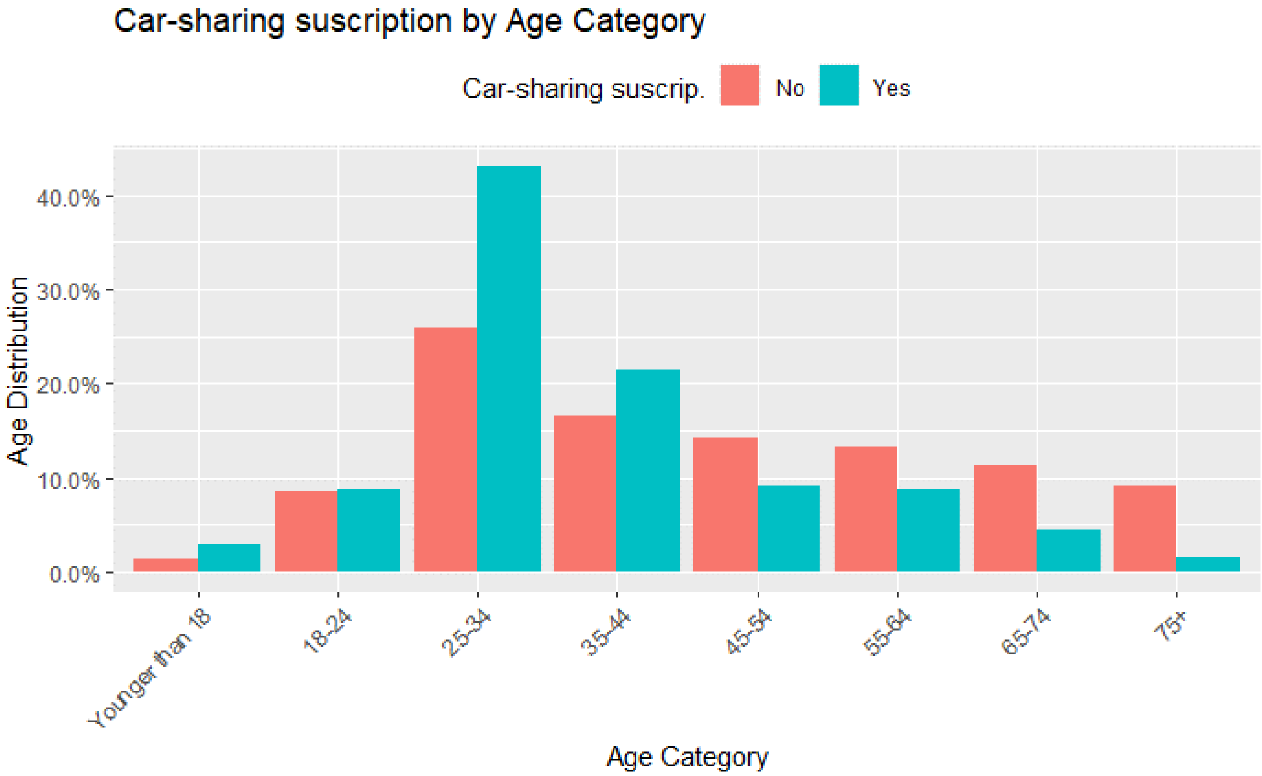


Figure 93. Age distribution of respondents from households with/without car-sharing subscriptions

3.2.2.2.3. Nationality

As for the effect that the nationality of the surveyed person may have on car-sharing subscriptions, the results show a very slight over-representation of Belgian respondents who do have at least one household subscription at home with respect to the control group (88.3% vs 87.9, respectively), but this does not seem to be a significant difference.

3.2.2.2.4. Labour status

Age differences are also reflected in the percentage of respondents in (in)active population: more than 80% of the surveyed people with household car-sharing subscriptions are part of the economically active population (employed or unemployed), compared to 60% in the control group.

Among the segments that make up the inactive population (students, homemakers, retirees and people with disabilities), student respondents are also slightly over-represented among the household car-sharing subscription group, while retired people are clearly under-represented compared to their proportion in the general population.

These results are consistent with those that can be found in the literature, such as the study conducted in Basel, Switzerland, which also observed a special over-representation of students and workers among car-sharing subscribers, especially self-employed workers (Becker, Ciari, & Axhausen, 2017).

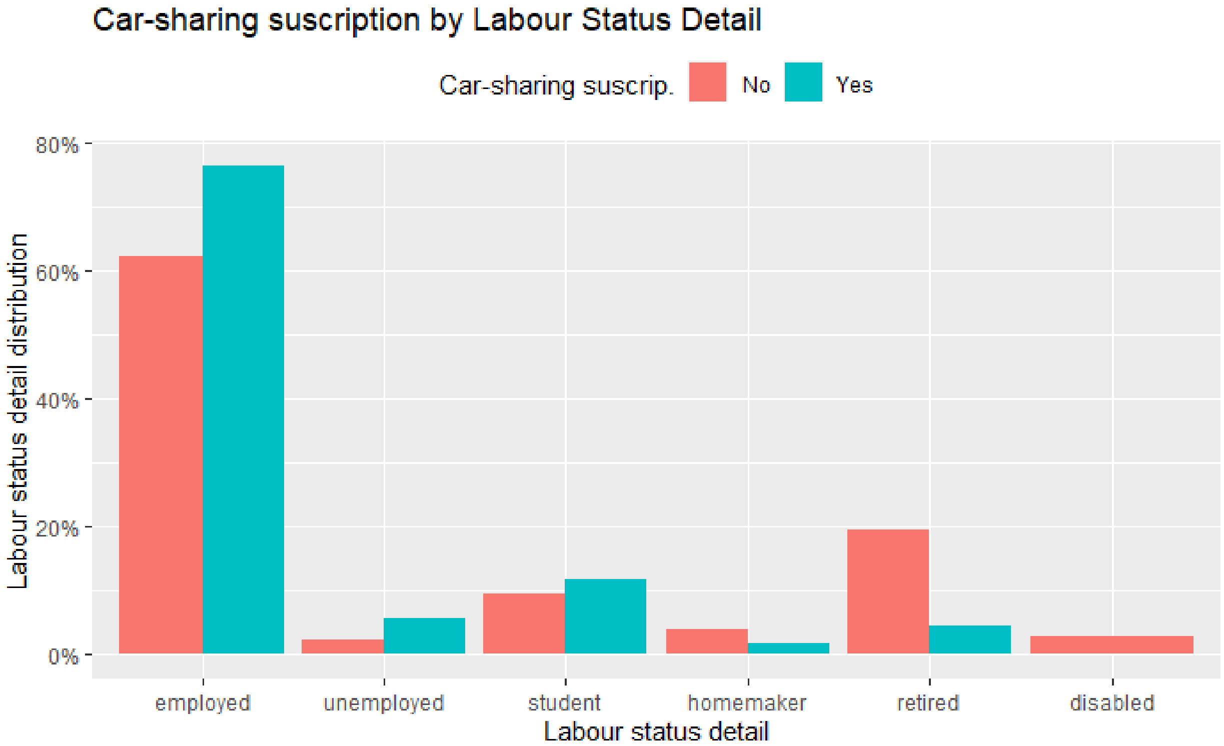


Figure 94. Labour status distribution of respondents from households with/without car-sharing subscriptions

The figure above also shows that no people with disabilities have been found among the household car-sharing subscription group. This probably reflects the lack of adaptation to this type of services for this sector of the population, which is undoubtedly an important aspect to take into account.

3.2.2.2.5. Education level

The survey data reflect that respondents living in a house with a least one car-sharing subscription are generally more educated than the rest of the population, with more than half of them having a Master's degree or equivalent. This observation is entirely consistent with that which can be found in the scientific literature, where, in general, more than 70% of car-sharing service subscribers have a Bachelor's degree or higher (R. Clewlow, 2016) (Becker, Ciari, & Axhausen, 2017) (Yoon, Cherry, & Jones, 2017).

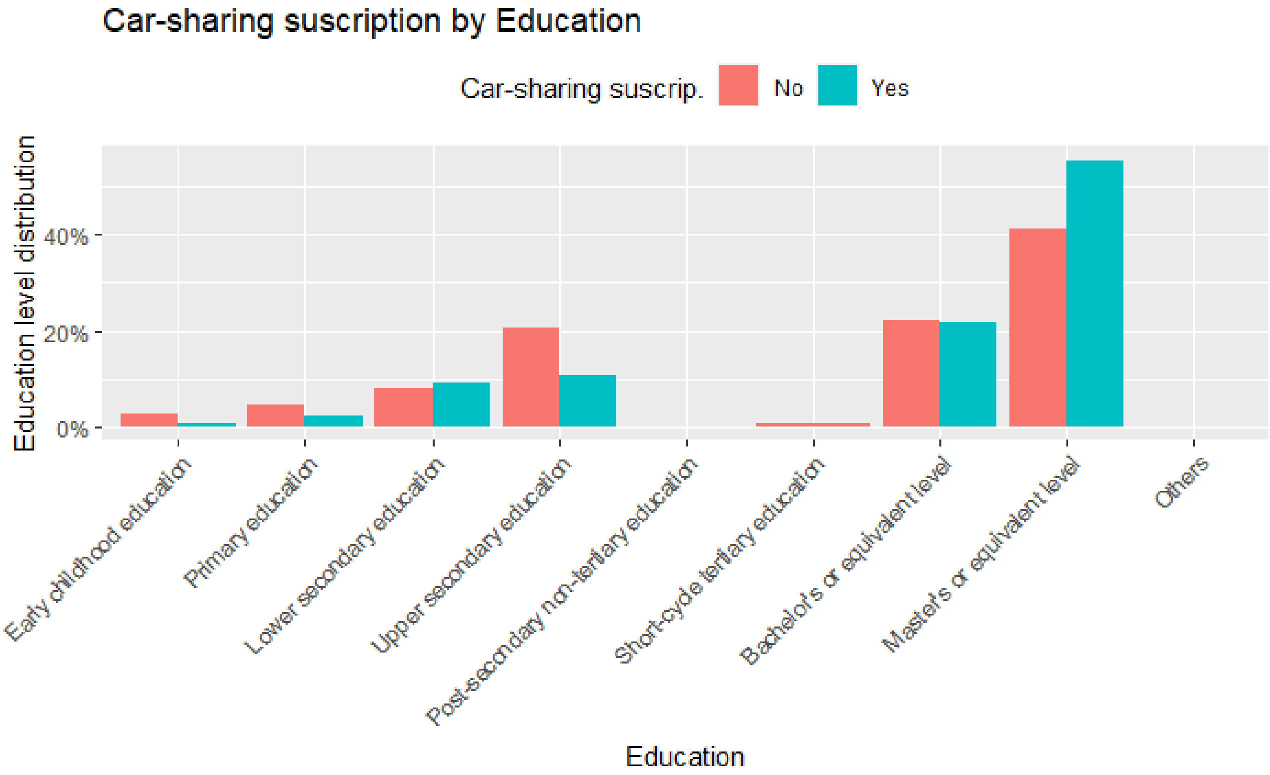


Figure 95. Education level distribution of respondents from households with/without car-sharing subscriptions

3.2.2.2.6. Incomes

As can be seen, households with at least one car-sharing subscription enjoy a higher average income than the rest of the population. This fact coincides with that reported in the literature (R. Clewlow, 2016) (Becker, Ciari, & Axhausen, 2017) (Yoon, Cherry, & Jones, 2017). In Leuven, one important exception is found in households whose income ranges between 1,000 and 1,499€ per month, where car-sharing subscribers vs non-car-sharing subscribers represent nearly the same percentage. Although the first intuition may lead to think that this is a sector of the population with a high number of students, and therefore that may be the reason for this balance, a more detailed analysis reveals that this is not the case.

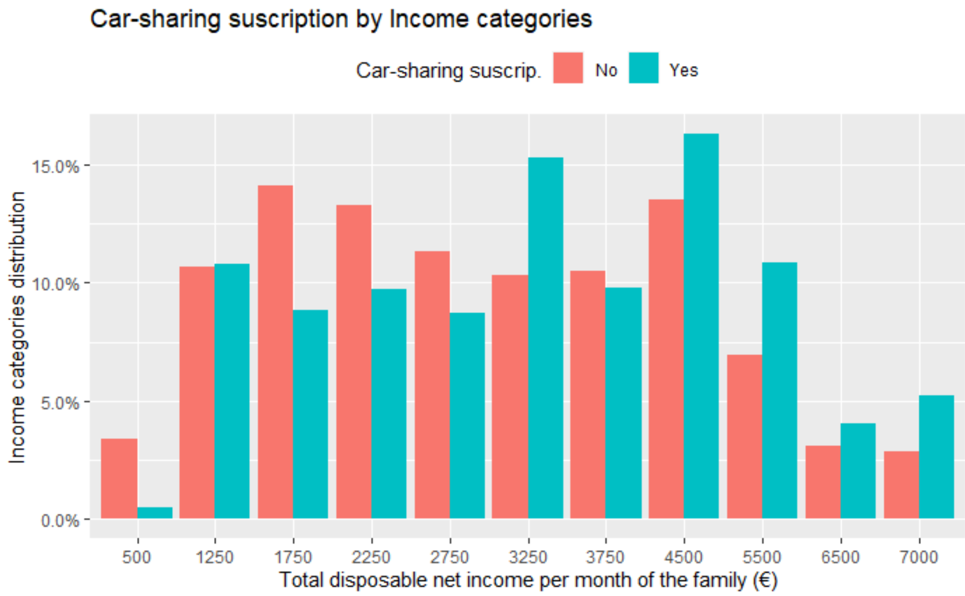


Figure 96. Incomes distribution of households with/without car-sharing subscriptions

In the following two graphs we show the distribution of labour status and type of household for individuals whose household is in the above income range (1,000 EUR-1,499 EUR). Focusing on the labour status, we see that individuals whose households have car-sharing subscriptions are mostly employed or unemployed, while in the control group they are retired, and that students are, in fact, a minority.

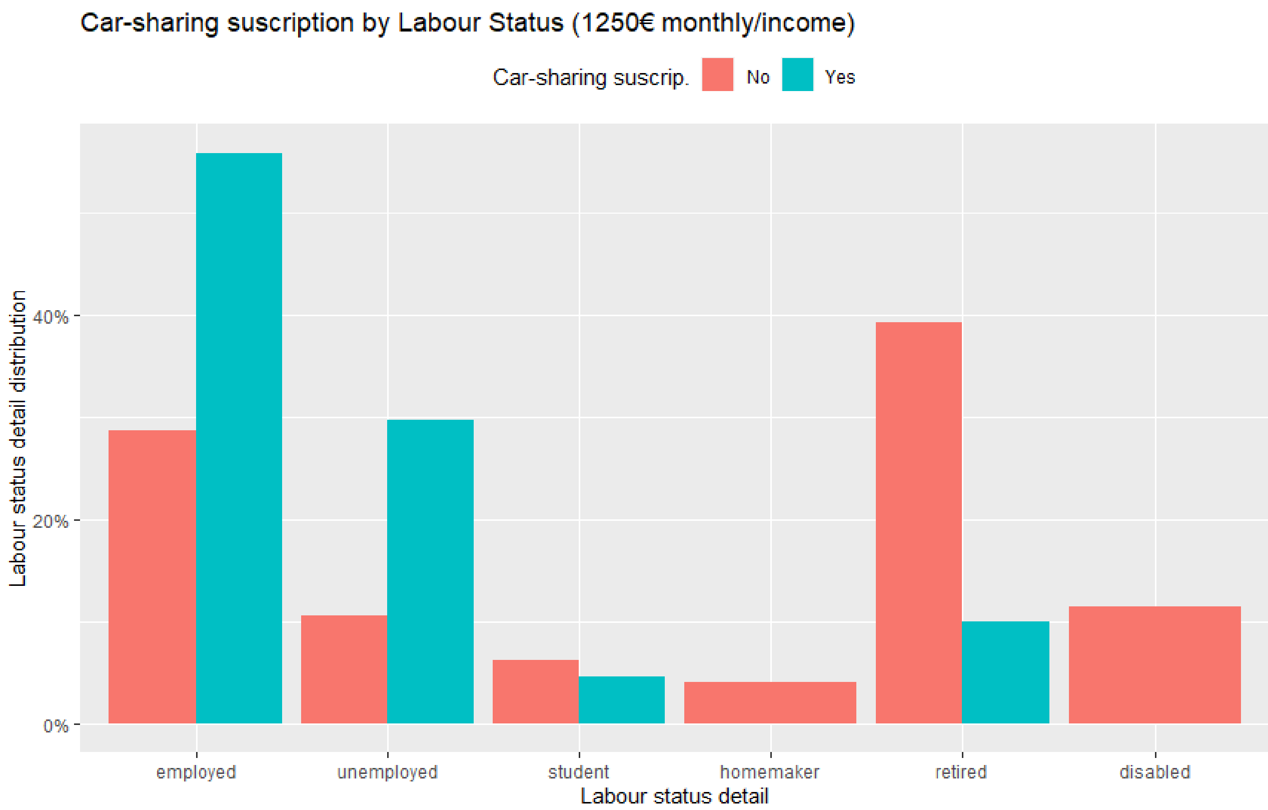


Figure 97. Labour status of respondents from households with/without car-sharing subscriptions with 1250€ monthly income

Looking at the type of household, although the category "single without children living at home" is by far the most common in both cases, individuals in households with car-sharing subscriptions are over-represented in the categories "single living with friends", "single with children living at home", "couple with children living at home" and "couple without children living at home". Of these four categories, the last three are particularly interesting, given the level of income. They refer to households with low purchasing power, probably because only one of the household members has a job. These results may be suggesting that car-sharing services are helping low-income households to access to the use of a car.

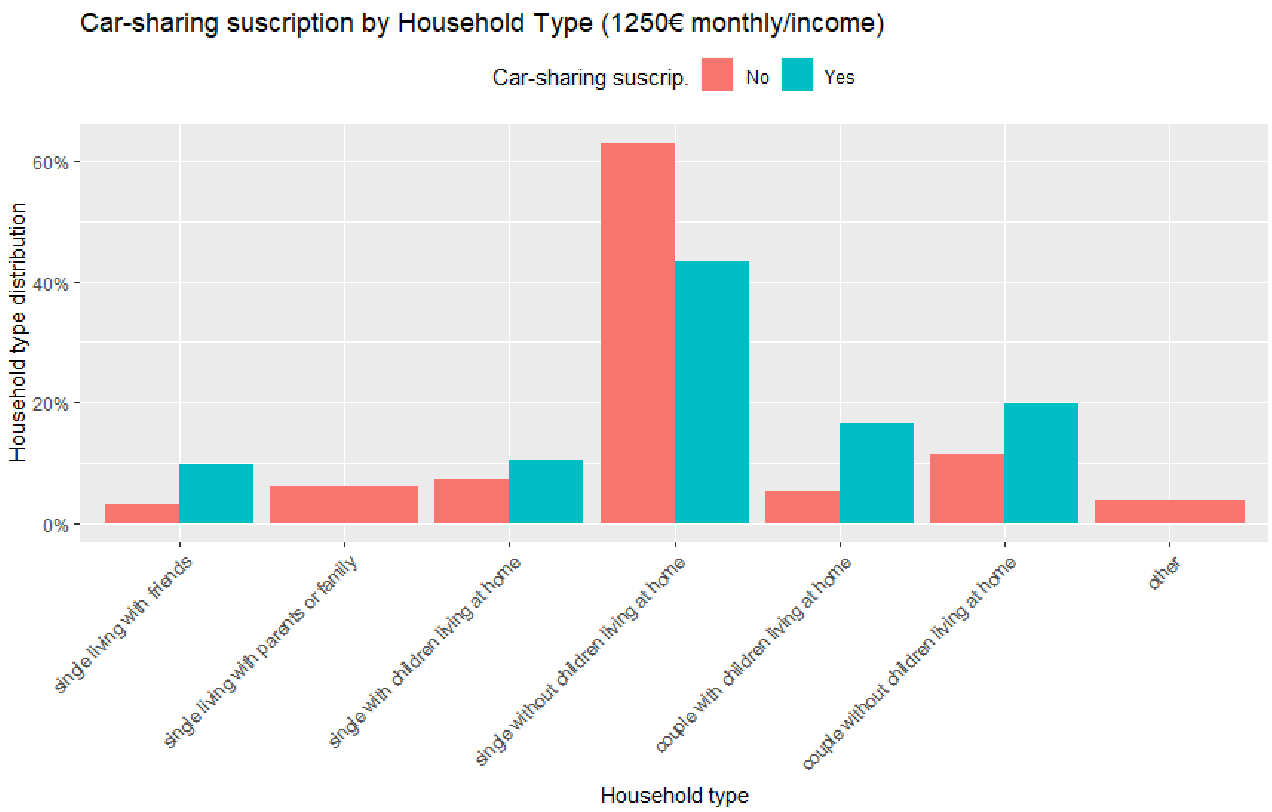


Figure 98. Household type distribution of households with/without car-sharing subscriptions with 1250€ monthly income

3.2.2.2.7. Household size and type

The average household size, in terms of people living under the same roof, is slightly larger in those homes that have at least one car-sharing subscription. This observation is compatible with results reported in the literature (Becker, Ciari, & Axhausen, 2017), although it is not usually a studied variable.

Figure 99. Household size distribution of households with/without car-sharing subscriptions

Relating the size of the household to the type of family, it is observed that the proportion of households with a subscription is lower in the case of singles with or without children. This is also in line with the previous plot where we saw that households with only one member tend to be less inclined to subscribe to car-sharing services. This seems to indicate that having several people living in the household may lead to a higher probability of subscribing to car-sharing services. Possible reasons for this are a higher economic capacity of households with more individuals or not having enough vehicles for all the members of that household.

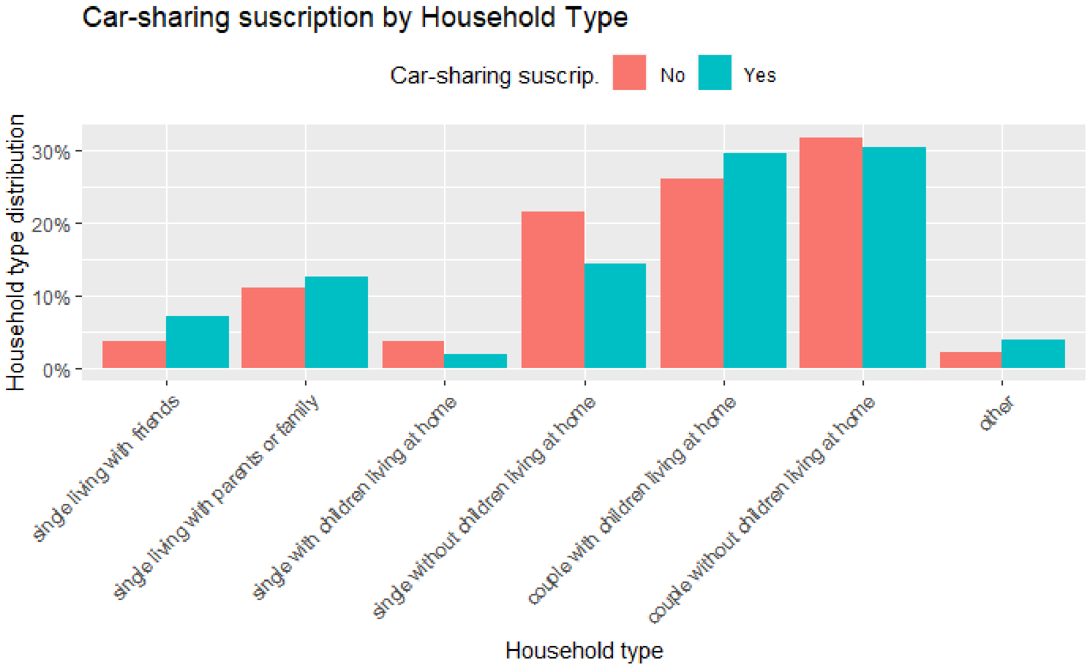


Figure 100. Household type distribution of households with/without car-sharing subscriptions

3.2.2.2.8. Main mode of transport

Looking at the main modes of transport to work or to the study centre, surveyed people within the household car-sharing subscription group stand out for their less frequent use of the car in favour of the bicycle and public transport (mainly train). This fact fits in with results reported in other works in the literature (R. Clewlow, 2016)(Yoon, Cherry, & Jones, 2017). These data also give indications that in Leuven, car-sharing services may be competing with public transport, although as we will see later with public transport passes, this competence may not be very high. However, this is a fact that needs to be validated, as the data are not conclusive. The exception to the previous statement is the category “on foot”, where there is a lower proportion of respondents with household car-sharing subscriptions who travel on foot. These results are not aligned with other studies such as one carried out in the San Francisco Bay Area (R. Clewlow, 2016), where results were obtained in the opposite direction. This is probably due to the different sizes and the urban planning characteristics of both cities.

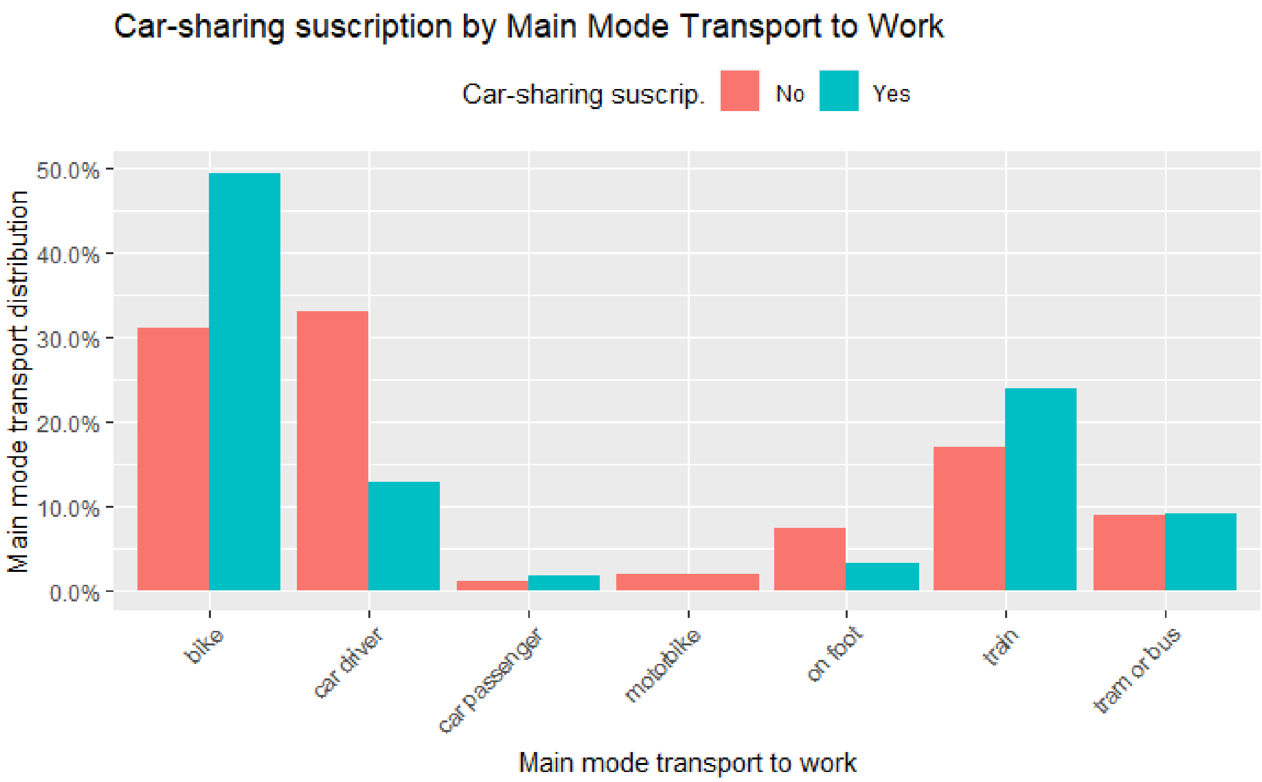


Figure 101. Main mode of transport distribution of respondents from households with/without car-sharing subscriptions

3.2.2.2.9. City districts

The results of the survey show that car-sharing members are concentrated in the centre of the city and the neighbourhoods closest to it (Leuven, Kessel-Lo and Heverlee Oost), while they drop significantly in other as areas (Heverlee West and Wilsele Wijgmaal). These could be connected to the supply and car-ownership in those city districts, as we will see in Section 3.2.2.2.13, although other factor could be also affecting.

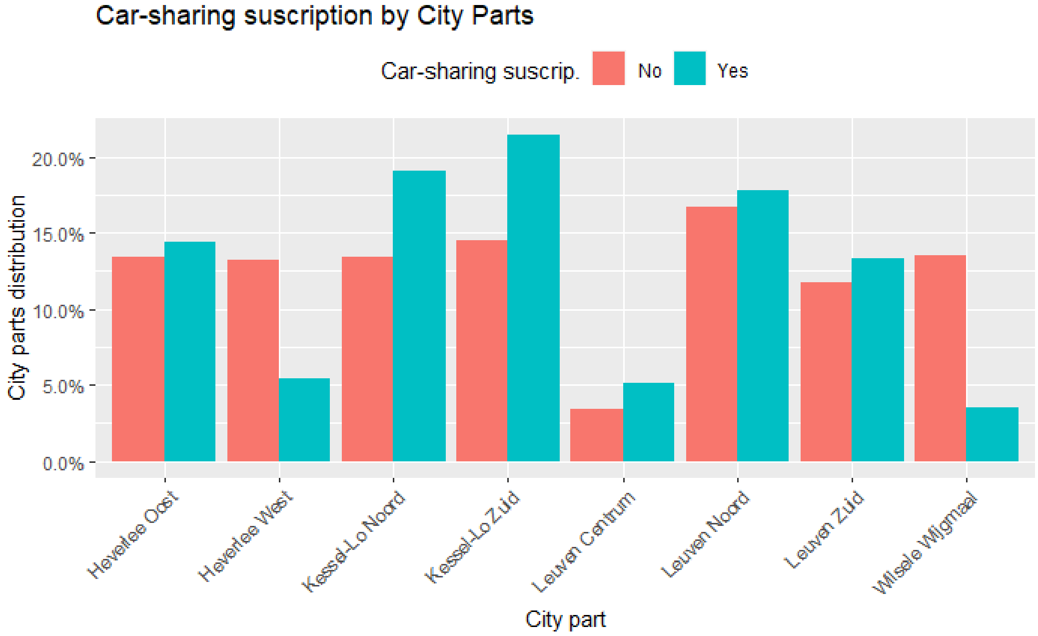


Figure 102. City part distribution of households with/without car-sharing subscriptions

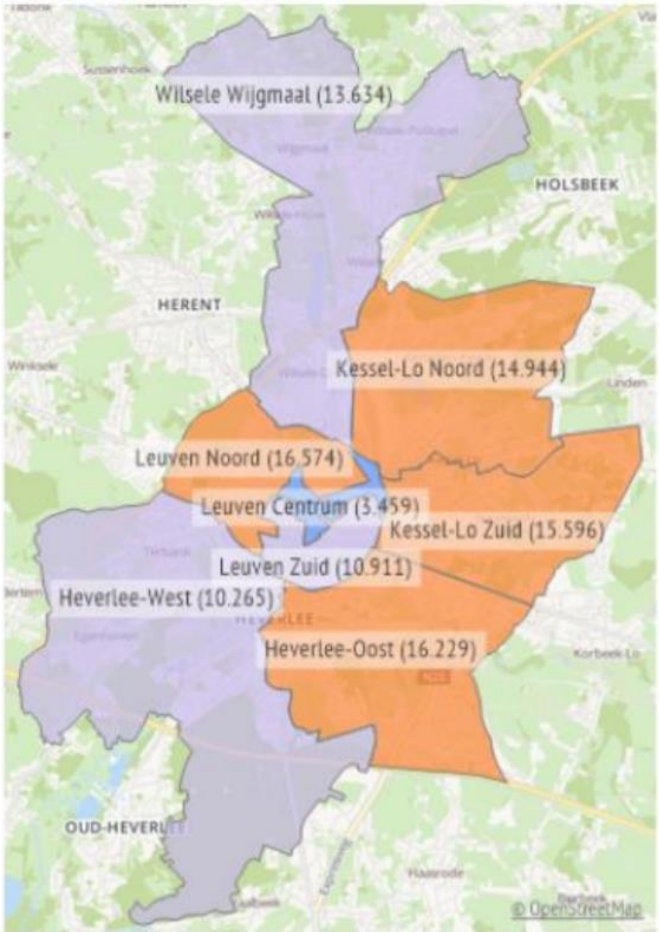


Figure 103. Leuven districts and population from 2019 national register [https://cwisdb.kuleuven.be/kaarten-bin/basiskaart.pl]

3.2.2.2.10. Satisfaction with city centre accessibility by public transport

In general, most of the population is satisfied with the degree of accessibility of the city centre offered by public transportation. However, a slightly bigger proportion of satisfied people is detected among those who have household car-sharing subscriptions, which reinforce the idea that these two transport modes may be competing.

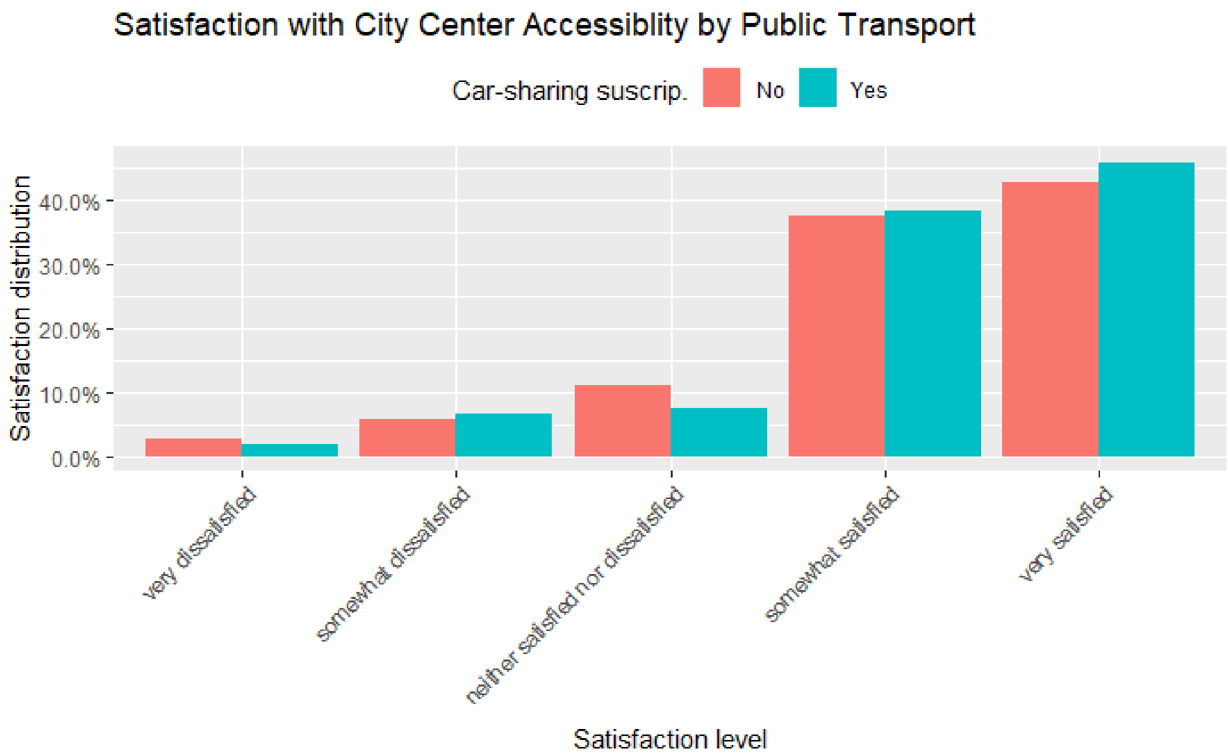


Figure 104. Distribution of satisfaction with the public transport of respondents from households with/without car-sharing subscriptions

3.2.2.2.11. Vehicle-Ownership

The survey results show that, on average, households with car-sharing subscriptions have significantly fewer cars: while households with a car sharing subscription are mostly over-represented among those having 0 cars of their own, the rest of cases the share is well below the base population. This fact has been clearly observed in other works in the literature (R. Clewlow, 2016)(Becker, Ciari, & Axhausen, 2017). It should be noted that more than 40% of households with car-sharing subscribers do not own any car.

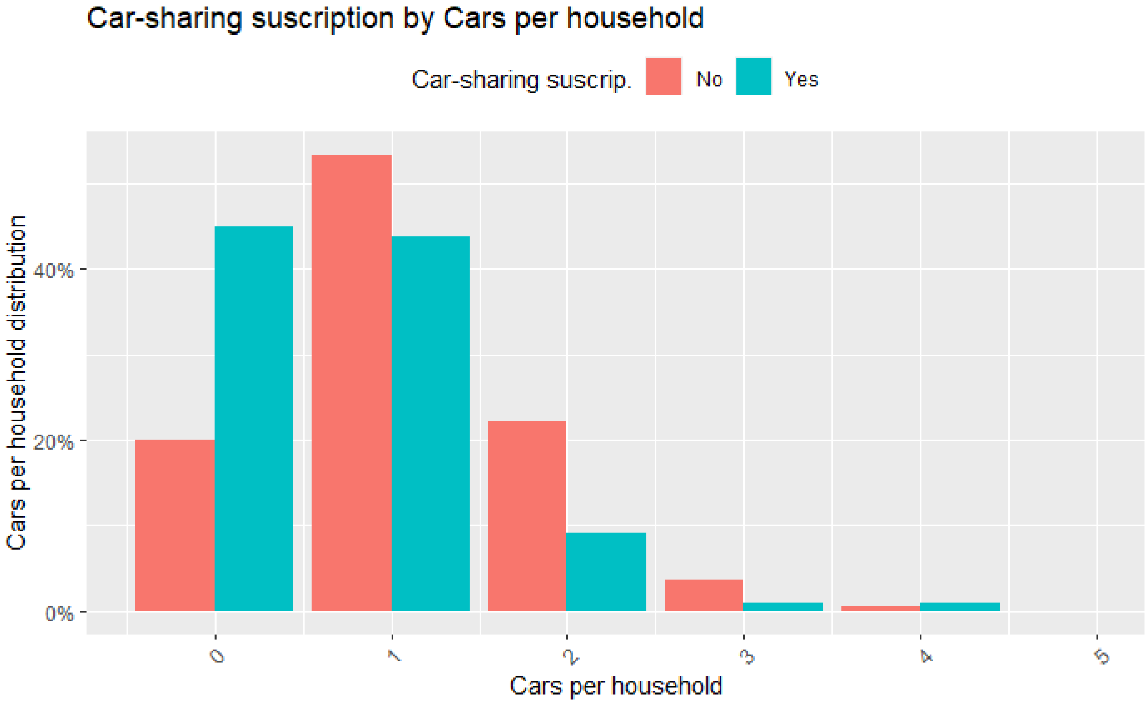


Figure 105. Car-ownership distribution of households with/without car-sharing subscriptions

Analysing motorcycles as an alternative motor vehicle option, households with car-sharing subscriptions also have fewer vehicles of this type, although the difference is very small. This fact has also been observed in other works (Becker, Ciari, & Axhausen, 2017).

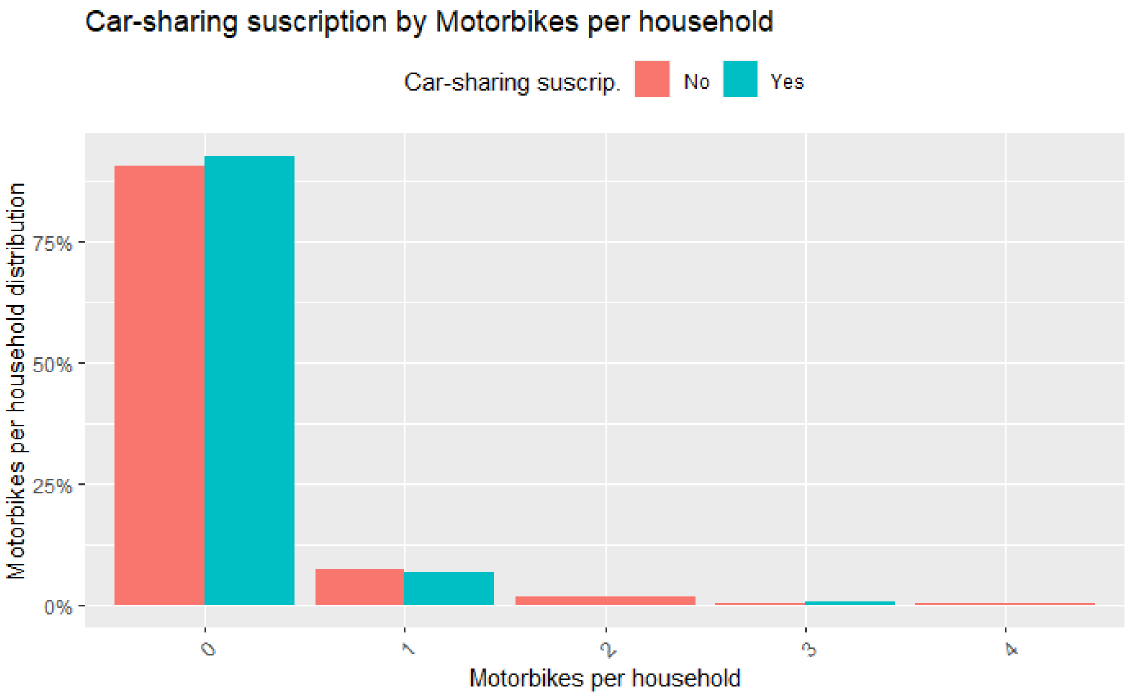


Figure 106. Motorbike-ownership distribution of households with/without car-sharing subscriptions

If we analyse bike-ownership, the plot below clearly shows that households with car-sharing subscriptions tend to own a higher number of bicycles, in line with what it is also observed in the literature (R. Clewlow, 2016) (Becker, Ciari, & Axhausen, 2017). Another possible reading of the graph is that households with a higher number of bicycles are more likely to have a car-sharing subscription, which again emphasises this complementarity between bicycle and car-sharing along with the mindset of their users.

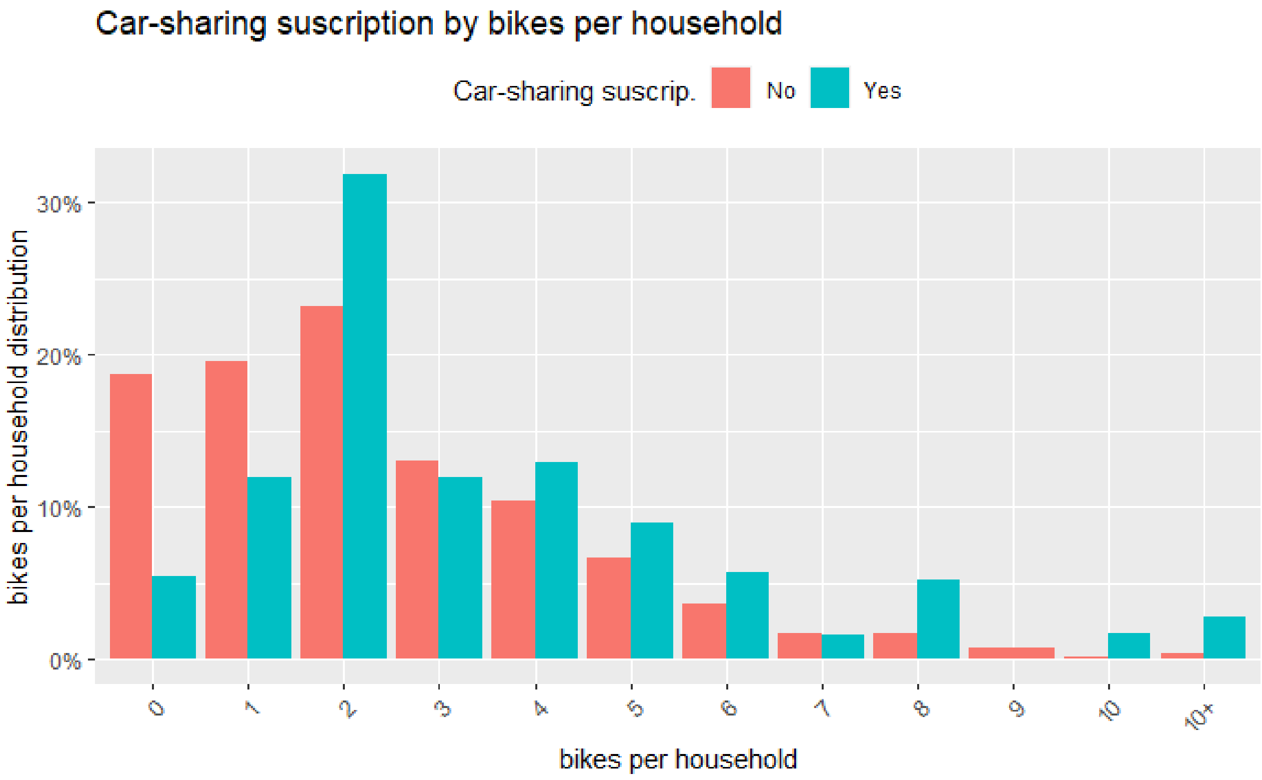


Figure 107. Bike-ownership distribution of households with/without car-sharing subscriptions

3.2.2.2.12. Public transport passes

In terms of the number of public transport passes, the survey results clearly show a higher average number of passes in those households that also have car-sharing subscriptions. This fact, together with a lower probability of having a car-sharing subscription when not having any kind of public transport pass, reinforces the idea that although there is competence between both modes of transport in Leuven, the negative impact of car-sharing in public transport is not high. Although here this trend is not so clear and the lack of information about the type of pass does not allow us to go further along this line.

Figure 108. Public transport passes distribution of households with/without car-sharing subscriptions

Although the survey does not distinguish between types of public transportation passes (annual, monthly, ...), there are papers in the literature that report the same observation, especially in terms of the number of annual passes (Becker, Ciari, & Axhausen, 2017).

3.2.2.2.13. Car-sharing supply

To conclude this sub-section, we are going to provide an analysis of the correlation between the supply of car-sharing and the subscription to this type of service. To this end, we have analysed the number of car-sharing vehicles distributed in each of the city districts described in Section 3.2.2.2.9, information that was provided by the city of Leuven, and on the other hand, the number of people living in households with at least one subscription to car-sharing services.

In the figure below we show a bubble chart where in the X axis we have the number of car-sharing vehicles available in each district and in the Y axis the number of people living in households with at least one subscription to car-sharing services. The colour of each bubble indicates to which district of the city it corresponds, while the size of the bubble is given by the population of that district. In addition, it shows in black the regression line that adjusts the variables indicated above for the X and Y axes, as well as the R^2 index of this regression.

The graph shows a positive correlation between the available car-sharing supply and the subscription to car-sharing services, but at the same time, it shows that this correlation is not perfect. Specifically, the R^2 index tells us that the supply of available car-sharing vehicles only explains 18% of the variance in the number of people living in households with at least one subscription to car-sharing services. The deviations of Wisele Wijgmaal and Kessel-Lo Zuid are interesting because despite having a similar population volume and car-sharing offer, the subscription to car-sharing services in both districts is very different. This indicates that other factors also play a significant role in the adoption of these services. To illustrate this, in Figure 110 we show the same chart, with the difference that the size of the bubbles indicates the number of people living in households without any private car, which as can be seen is abnormally low in Wisele Wijgmaal, which may in part explain the low adoption of car-sharing services despite the high supply in that district. In fact, that figure also shows the R^2 index that results from fitting a linear model as follows: $y=ax+bz+c$, where y would be the number of people living in households with at least one car-sharing subscription, z would be the number of people living in households without any private car, while a, b, and c are coefficients. As can be seen, this new variable allows an improved fitting since the R^2 index goes from 0.18 to 0.36, but both factors still have no great explanatory power.

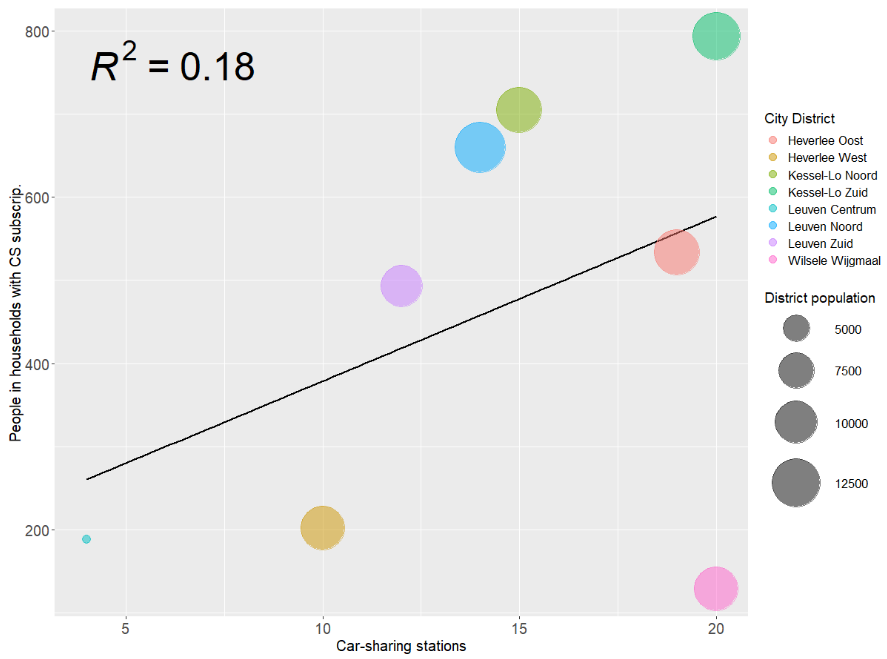


Figure 109. Correlation between car-sharing supply and people in households with/without car-sharing subscriptions (District Population)

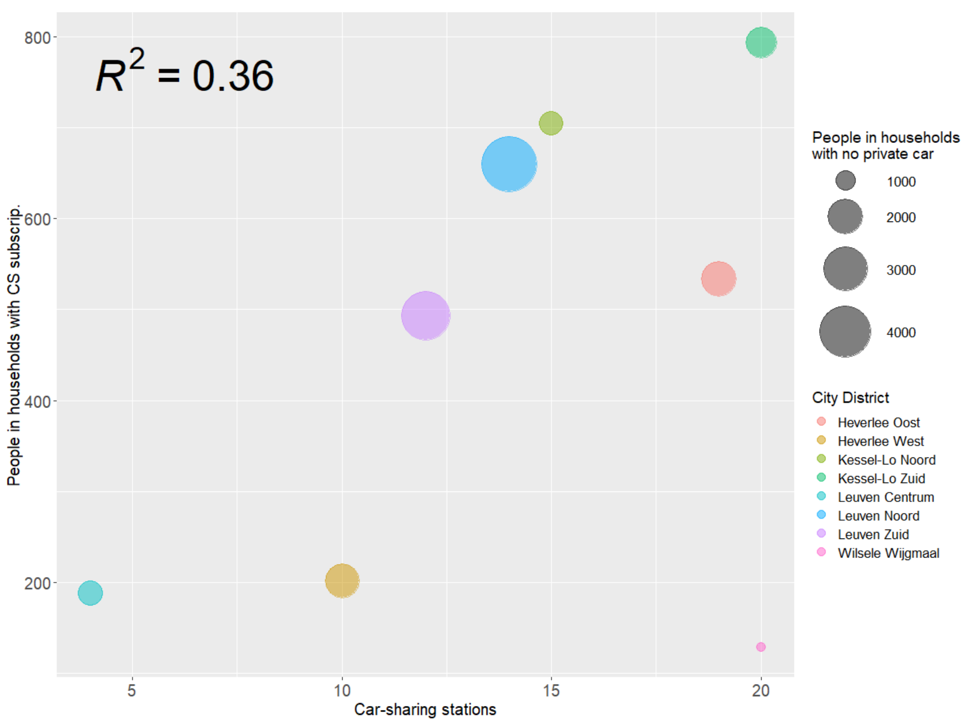


Figure 110. Correlation among car-sharing supply and people living in households with no private car with people living in households with at least one car-sharing subscription

3.2.2.3. Analysis of most relevant factors for car-sharing subscription using white-box Machine Learning

In the previous analysis, we studied the influence that different factors such as sex, education, work status, etc. have on the probability that a person lives in a household where there is at least one subscription to a car-sharing service, which in turn can be interpreted as a higher frequency of use of these services.

In order to identify which factors have a higher discriminatory power in determining the probability of a person living in a household where there is at least one car-sharing subscription, we have used White-box Machine Learning techniques for this purpose, and more specifically Decision Trees. These techniques automatically extract rules that allow us to infer the value of a dependent variable Y from the independent variables X1,...,Xn. In this case, the dependent variable is a binary variable that indicates whether the person lives in a household with at least one car-sharing subscription (CS Subscription) or not (No CS Subscription). The independent variables are all the factors analysed in the previous section.

In order to build the decision tree, it has been necessary to use Machine Learning techniques for unbalanced classification problems (Kaur,Pannu, & Malhi, 2020), since the difference between the number of people in households with car-sharing subscriptions versus those who do not is very high, namely, 3,704 versus 78,180, or in other words, 4.7% versus 95.3%. Specifically, we have used techniques called "cost-sensitive" which consist of weighting each sample by the inverse of the total number of samples in the class to which the sample in question belongs so that an example of the minority class (CS Subscription) has a greater weight than in the case of the majority class (No CS Subscription). In this way, getting one example of the minority class (CS Subscription) has a greater weight than in one of the majority class (No CS Subscription). That is, all those samples that correspond to CS Subscription have been given a weight of 1/3,704, while the rest have been given a weight of 1/78,180.

For the construction of the decision tree model, we have used the Rpart library of R[8] with the default parameters, except for the complexity parameter (cp) and the parameter that controls the minimum number of observations that must exist in a node in order for a split to be attempted (minsplit). Their values were set to 0.03 and 200, respectively. The objective was to control the depth of the tree to make it interpretable and, at the same time, to ensure that leaf nodes represent a significant number of samples, to avoid a possible over-fitting that could lead to erroneous conclusions.

The decision tree obtained can be seen in the following figure. The blue nodes are decision nodes that have a condition associated with them. The blue box indicates the variable involved in the condition, while the white box indicates the condition. The branches that go to the left correspond to the fulfilment of that condition, and the branches to the right the opposite. The orange and green nodes are leaf nodes, which indicate for which of the two possible values of the dependent variable its probability increases with respect to the a priori probability (95.3% and 4.7%, for No CS Subscription and CS Subscription, respectively). In the white box of the leaf nodes, Prob NCSS/CSS indicates both probabilities of No CS Subscription and CS Subscription, respectively, while N indicates the number of individuals who meet those conditions according to the survey scaling factors.

[8] https://www.rdocumentation.org/packages/rpart/versions/4.1-15/topics/rpart

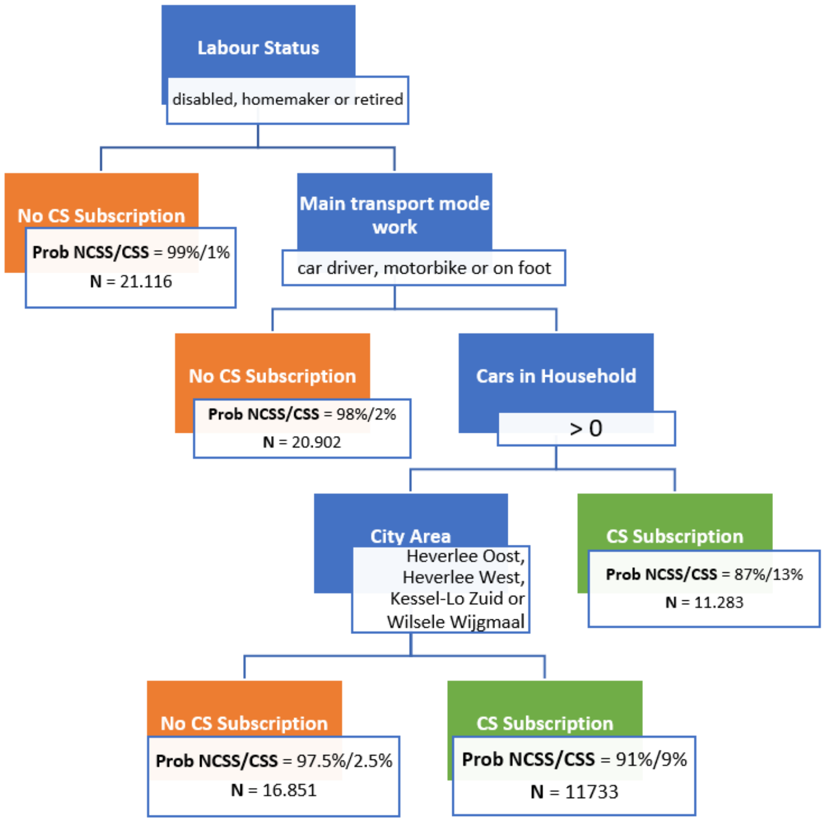


Figure 111. Decision Tree with the most relevant factors for car-sharing subscription

Looking at the figure, we see that one of the most relevant variables for discriminating between subscription and non-subscription in the household is employment status. Specifically, if a person is disabled, homemaker or retired, the probability of having a subscription in the household is reduced by more than a quarter. The next most relevant variable of those studied is the main transport mode used to get to work. The employed, student or unemployed who usually go to work by car (as a driver), motorbike or on foot has only a two per cent probability of having a subscription in their home. When this condition is not met, the most discriminatory variable is the number of vehicles in the household. The reading to be made is that employees, students or unemployed people who usually go to work by bicycle or public transport (train, bus, tram) and who do not have a car at home, multiply by almost three the probability of having a car-sharing subscription at home. Finally, another variable that has been shown to be relevant is the city district in which the household is located. Thus, if all the conditions indicated in the tree are met in terms of work status, transport mode to work and the number of cars at home, if you live in the areas Heverlee Oost, Heverlee West, Kessel-Lo Zuid, Wilsele Wijgmaal the probability of having a car-sharing subscription at home is almost halved. In contrast, if you live in the Centrum, Kessel-Lo Noord, Leuven Zuid and Noord areas, you are twice as likely. This is probably due to factors such as age, purchasing power and the type of households in those areas.

3.2.2.4. Profiling of car-sharing subscribers in Leuven

The objective of this subsection is similar to the previous one in that it aims to study the interrelations between the different variables. However, this time the study is carried out by segmenting those users in whose households there is at least one subscription to car-sharing services. In this way, we aim to build a profile of the different types of individuals within this group based on characteristics such as their sex, age, income, education, household size and type, as well as the number of bicycles, cars and public transport passes in the household.

To carry out this segmentation we use unsupervised learning techniques and specifically the k-means. Given that some of the variables analysed were categorical, we opted to create binary dummy variables, to make the application of this method possible. For those variables that were ordinal, it was chosen to use a numerical scale (e.g. 1,2,3...). Once these transformations were made, a normalization of the variables was applied so that all of them were on the scale [0,1]. To decide the appropriate number of clusters, that is, the value of the parameter k, we used the elbow rule, which is one of the most common approaches for this purpose. After the application of this rule, the value of the parameter k was set to four. The sample used had a size of 110 elements but since k-means algorithm cannot deal with missing values, 16 samples need to be discarded. In this way, we had a total of 96 samples accounting for 3084 individuals when the expansion factor is applied.

Tables Table 31, Table 32, Table 33 and Table 34 show the profile for each of the variables considered for the four clusters or profiles obtained. Some of these variables are categorical and their numerical coding is presented in Table 30.

Numerical coding	Categorical variables		
	Sex	**Age**	**Education**
1	Male	Younger than 18	Early childhood education
2	Female	18-24	Primary education
3		25-34	Lower secondary education
4		35-44	Upper secondary education
5		45-54	Post-secondary non-tertiary education
6		55-64	Short-cycle tertiary education
7		65-74	Bachelor's or equivalent level
8		75+	Master's or equivalent level

Table 30. Coding used for categorical variables in car-sharing subscriber profiling

The first profile we have, shown in Table 31, are young couples, aged between 25 and 44, with children. These are employed people who have a high level of education and high purchasing power. They live in households with an average size of about four people, which can be interpreted as having an average of two children. A remarkable feature is that more than half of these households have at most one car and the 75% have 5 or more bicycles at home.

Table 32 shows the second profile. In this case, we also have young couples but aged between 18 and 34, and without children, so the average household size is 2. They are mostly employed people with an intermediate income level and a very high level of education. In this case, 75% of the households have at most one car and at least five bikes.

The third group corresponds to single people without children with an average age between 25 and 44 years, with a slightly higher number of men. They have a medium-high educational level and the majority work (68%), although there is also an important portion of retired people (26%). They live in medium-low income households, and they do not own a car. The number of bicycles at home is also high (75% have at least two bicycles).

Finally, in the fourth group, the profile we find is that of single people (with a high majority of men) living with parents or with friends. They are mostly young, but there are also older people, probably related to those households in which single lives with parents. They are also highly educated, mostly employed (68%) but also students (25%) and have a high purchasing power. As in the previous cases the number of private cars in the household is low (50% has at most one car) and the number of bikes high (the mean is 2 bicycles per household member).

	Sex	Age Cat	Educ isced	Income	Labour Status	Household Size	Household Type	Household Bikes	Household Cars	Household PT passes
					Employed		Couple with children living at home			
Mean	1.50	3.46	7.11	4008.93	0.93	4.25	1.00	7.04	0.86	2.43
Std.Dev	0.51	0.92	1.57	1556.77	0.26	1.00	0.00	2.36	0.76	1.83
Min	1.00	2.00	3.00	1250.00	0.00	3.00	1.00	1.00	0.00	1.00
Q1	1.00	3.00	7.00	2750.00	1.00	3.50	1.00	5.50	0.00	1.00
Median	1.50	3.00	8.00	3750.00	1.00	4.00	1.00	7.00	1.00	2.00
Q3	2.00	4.00	8.00	5500.00	1.00	5.00	1.00	8.50	1.00	3.50
Max	2.00	5.00	8.00	7000.00	1.00	6.00	1.00	11.00	2.00	6.00

Table 31. Descriptive statistics for Group 1 (N = 936 individuals)

	Sex	Age Cat	Educ isced	Income	Labour Status	Household Size	Household Type	Household Bikes	Household Cars	Household PT passes
					Employed		Couple without children living at home			
Mean	1.35	2.77	7.29	3532.26	0.81	2.06	1.00	4.55	0.58	2.32
Std.Dev	0.49	1.31	1.44	1143.35	0.40	0.25	0.00	1.79	0.62	1.49
Min	1.00	2.00	3.00	1250.00	0.00	2.00	1.00	1.00	0.00	1.00
Q1	1.00	2.00	7.00	2750.00	1.00	2.00	1.00	5.00	0.00	1.00
Median	1.00	2.00	8.00	3250.00	1.00	2.00	1.00	5.00	1.00	2.00
Q3	2.00	3.00	8.00	4500.00	1.00	2.00	1.00	5.00	1.00	2.00
Max	2.00	7.00	8.00	5500.00	1.00	3.00	1.00	7.00	2.00	6.00

Table 32. Descriptive statistics for Group 2 (N = 1039 individuals)

	Sex	Age Cat	Educ isced	Income	Labour Status		Household Size	Household Type	Household Bikes	Household Cars	Household PT passes
					Employed	Retired		Single without children living at home			
Mean	1.42	3.84	6.79	1973.68	0.68	0.26	1.05	0.95	3.32	0.00	1.84
Std.Dev	0.51	1.77	1.81	776.81	0.48	0.45	0.23	0.23	2.00	0.00	0.90
Min	1.00	2.00	3.00	500.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00
Q1	1.00	2.00	7.00	1250.00	0.00	0.00	1.00	1.00	2.00	0.00	1.00
Median	1.00	4.00	8.00	1750.00	1.00	0.00	1.00	1.00	2.00	0.00	2.00
Q3	2.00	6.00	8.00	2250.00	1.00	1.00	1.00	1.00	5.00	0.00	2.00
Max	2.00	6.00	8.00	3750.00	1.00	1.00	2.00	1.00	7.00	0.00	5.00

Table 33. Descriptive statistics for Group 3 (N = 480 individuals)

	Sex	Age Cat	Educ isced	Income	Labour Status		Household Size	Household Type			Household Bikes	Household Cars	Household PT passes
					Employed	Student		Single living with parents or family	Single living with friends	Other			
Mean	1.19	3.12	6.38	4109.38	0.69	0.25	4.44	0.38	0.31	0.25	7.69	1.19	3.06
Std.Dev	0.40	2.28	2.09	2087.60	0.48	0.45	2.13	0.50	0.48	0.45	2.36	1.11	2.35
Min	1.00	1.00	2.00	1250.00	0.00	0.00	2.00	0.00	0.00	0.00	4.00	0.00	1.00
Q1	1.00	2.00	5.50	2250.00	0.00	0.00	3.00	0.00	0.00	0.00	5.50	0.50	1.00
Median	1.00	2.00	7.00	3750.00	1.00	0.00	4.00	0.00	0.00	0.00	8.00	1.00	1.50
Q3	1.00	4.50	8.00	6000.00	1.00	0.50	6.00	1.00	1.00	0.50	9.50	1.50	5.00
Max	2.00	8.00	8.00	7000.00	1.00	1.00	10.00	1.00	1.00	1.00	11.00	4.00	7.00

Table 34. Descriptive statistics for Group 4 (N = 629 individuals)

3.2.3. Characterization of car-sharing willingness

3.2.3.1. Applied methodologies

To try to characterize the people who show a greater predisposition to use car-sharing services, this analysis is based on exploiting the responses obtained by the City Monitor survey to the question “Do you already do car-sharing or are you willing to do it in future?” (Question 59 in the original survey), where possible answers are:

- I already do it
- I want to do it
- I won’t do it

Analysing these three groups under different socio-demographic, economic and transport habit variables, the aim is to profile the characteristics that best represent the willingness or unwillingness to do car-sharing.

As in the previous analysis, the number of interviewed individuals is 2,669, which using the expansion factor amounts to a total of 79,574 people. The distribution of answers in the three groups is the following: already doing it (N=9,479), want to do it (N=19,441), will not do it (N=50,665).

3.2.3.2. Result analysis

3.2.3.2.1. Sex

Consistent with the previous section, there is an over-representation of women who already use car-sharing services, while the other two options (to do it in the future or not) are balanced in terms of the sex of the person. As also discussed in the previous section, this result is not consistent with current literature where men are usually the over-represented group. This makes Leuven an interesting exception that could help to understand how the use of car-sharing services can be increased among the female population.

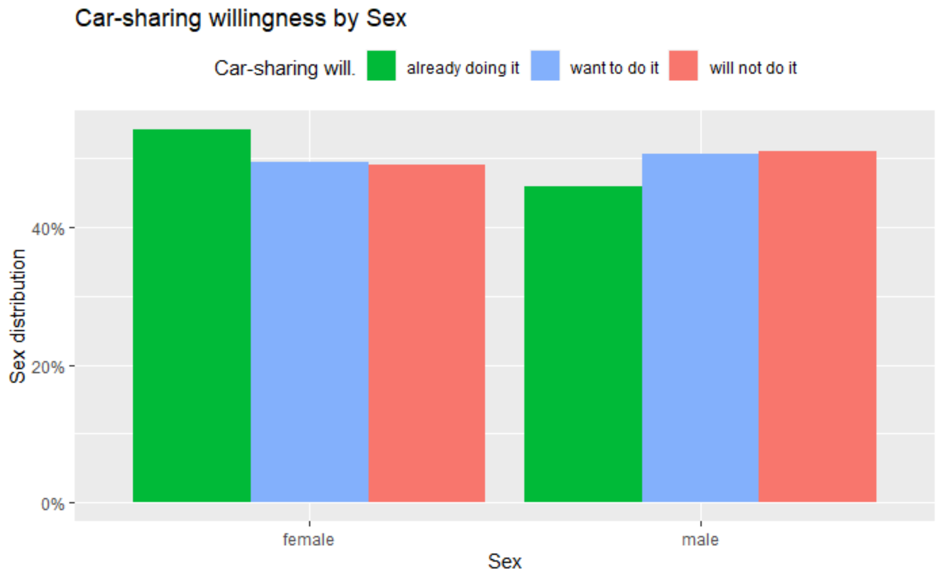


Figure 112. Sex distribution of car-sharing willingness categories

3.2.3.2.2. Age

As can be seen in the following figure, the survey data clearly reflect the influence of the generational component in the car-sharing service: in the under-34 age groups, there is a clear over-representation of people who already use or want to use car-sharing services; this attitude is progressively attenuated in the 35-54 age groups, although there is still some intention to start using such services; in contrast, from the age of 55 onwards, the proportion that does not contemplate using car-sharing services is clearly increasing.

These results fit in with those observed in the previous section, corresponding to the profile of car-sharing subscribers.

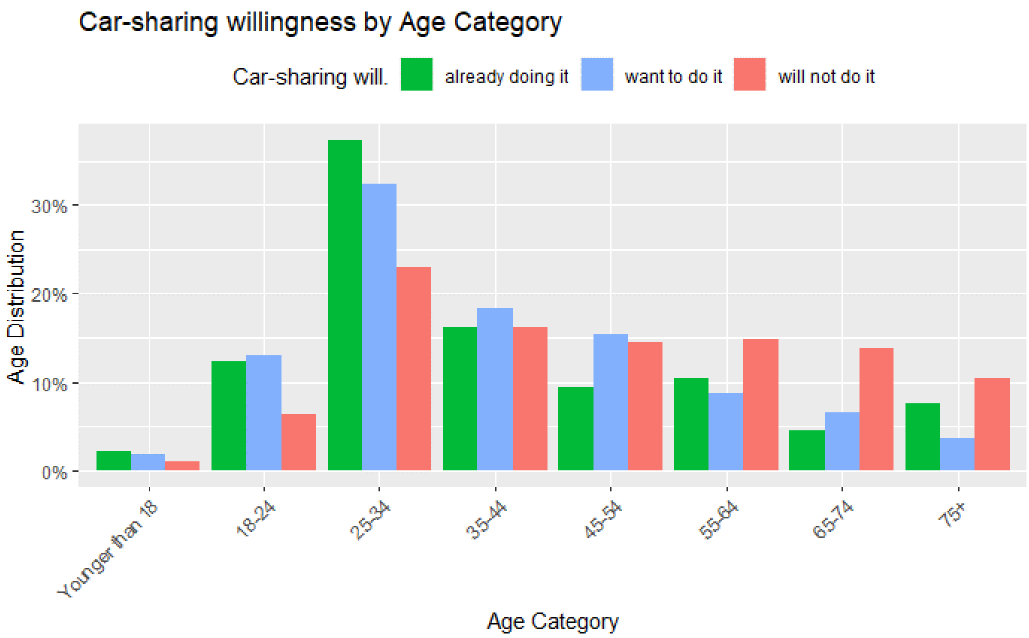


Figure 113. Age distribution of car-sharing willingness categories

3.2.3.2.3. Nationality

In terms of the nationality of the interviewees, it is observed that non-Belgians show a more receptive attitude towards the car-sharing service, with higher proportions among the groups that are currently using it (18%) or want to do it (15%) compared to their proportion in the total population (12%). Logically, the population of Belgian people appears over-represented in the group of people who do not intend to use these services (concretely, 91% when they are 88% of the population).

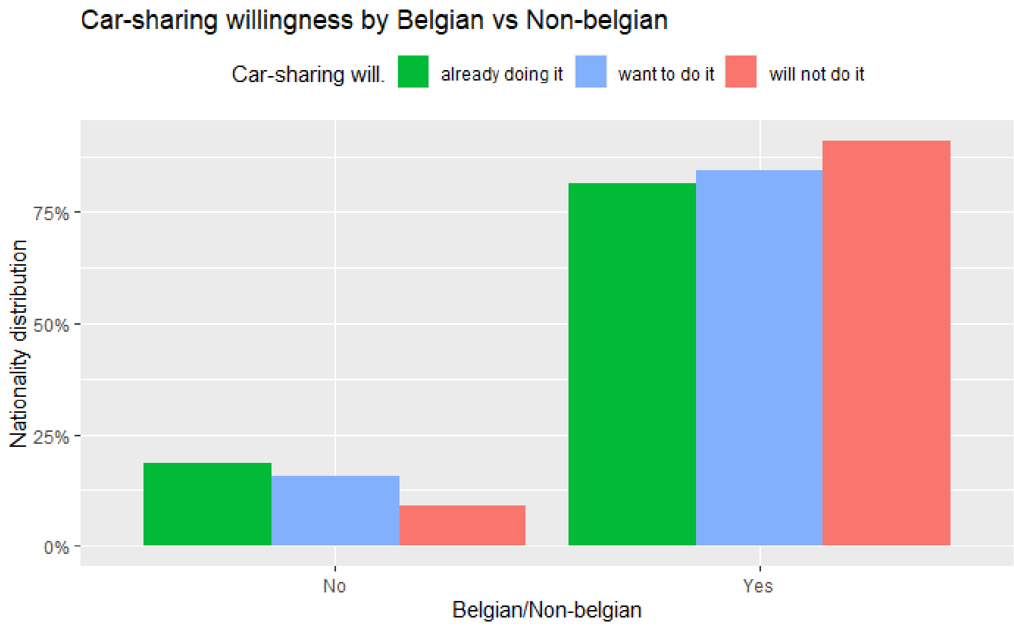


Figure 114. Nationality distribution of car-sharing willingness categories

These data differ from those seen in the car-sharing subscription analysis, where the Belgian population was slightly over-represented. The reason may be due to an artefact created in the analogous analysis in the previous section by considering people living in households with a carsharing subscription and not the subscribers themselves. Given that the Belgian population is largely in the majority, the share of Belgians actually related to a car-sharing subscription is probably being overestimated in the analogous previous analysis.

3.2.3.2.4. Labour status

Analysing in more detail the labour occupation, as shown in the following figure, where we can see higher over-representation towards doing car-sharing or wanting to do it, with respect to its share in the general population, are the employed, unemployed and students. Here it is important to highlight that the last two present a higher difference with the “will not do it” group. On the opposite side, we retired people where the group that does not want to do car sharing present a much higher share, confirming again the strong generational component on car-sharing. Unlike the analogous analysis in the previous section, we can see that disabled people are doing car-sharing (probably in a less regular way) and that there is a slightly higher share of disabled people in the “want to do it” group.

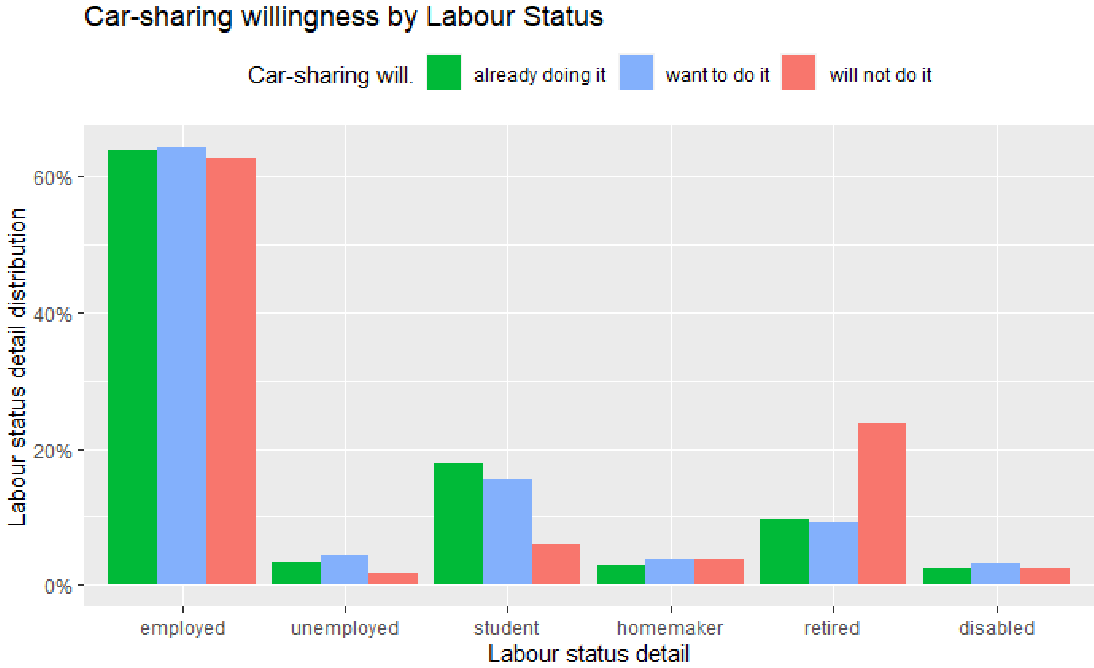


Figure 115. Labour status distribution of car-sharing willingness categories

3.2.3.2.5. Education level

Similar to car-sharing subscribers, the attitude towards such service is clearly influenced by education level, with the group with Master's study or equivalent being the most willing to use it. On the other hand, the group with a Bachelor's degree or equivalent is balanced with respect to its proportion in the total population, while in the rest of the education levels the rejection to this transport mode tends to be dominant.

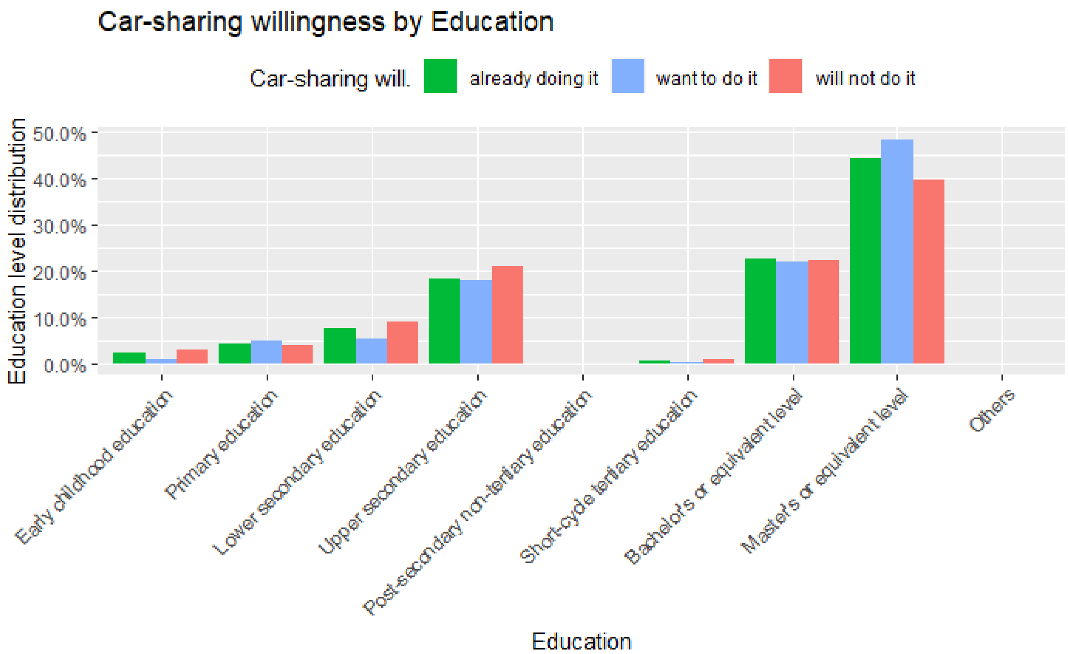


Figure 116. Education level distribution of car-sharing willingness categories

3.2.3.2.6. Incomes

Unlike what we saw earlier in the section on car-sharing subscribers, which they had higher average incomes when we classify the population towards their willingness to use this service, we see that the income distribution of people that are already doing car-sharing is biased towards lower incomes when compared to people that will not do it. One possible reason for this is perhaps the frequency of use, which we cannot observe in this survey. People who use these services more frequently (e.g. have a subscription) do have higher incomes, but those who use them less frequently or sporadically tend to have lower incomes. Also interesting is the fact that people who are doing car-sharing are over-represented in the two lowest income groups (500€-999€ and 1,000€-1,499€) and in two of the highest income groups (5,000€-5,999€ and >7,000€). This was also observed, although not as clearly, in the analysis of car-sharing subscription. Finally, if we look at the "want to do it" group, we see that it is somewhere in between the "already doing it" and "will not do it" groups.

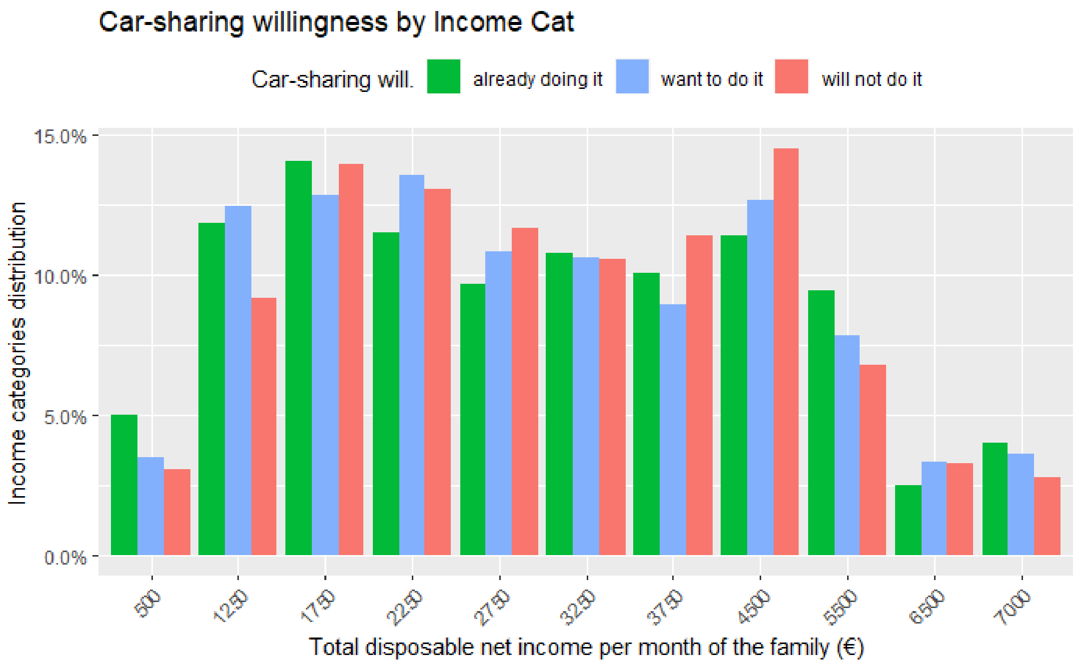


Figure 117. Incomes distribution of car-sharing willingness categories

3.2.3.2.7. Household size and type

With respect to household size in terms of how many people live under the same roof, the survey data show that, on the one hand, households with 5 or more inhabitants have a higher proportion of people using or willing to use car-sharing than they do in the total population. And, on the other hand, households with only two people represent a higher proportion of the group of people who do not contemplate using car-sharing. This fact may be related to couples without children or with children that do not live at home.

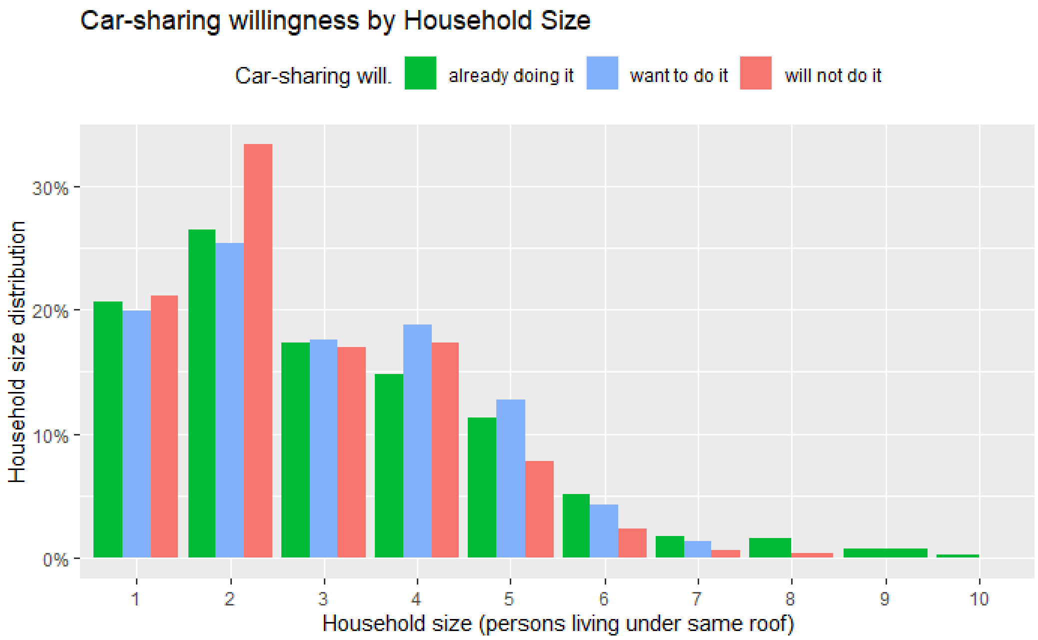


Figure 118. Household size distribution of car-sharing willingness categories

Actually, classifying households by type, we can see that the data support this possible explanation. In addition, it is also observed that the houses where singles live with friends or family (therefore, with a greater number of inhabitants), have a greater tendency to use car-sharing.

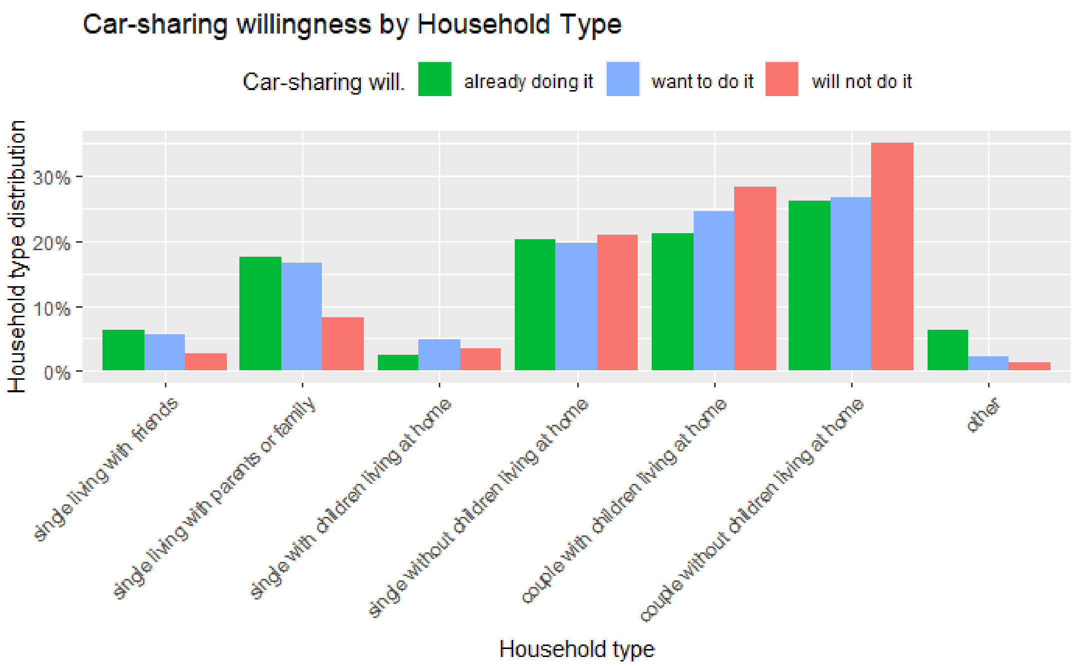


Figure 119. Household type distribution of car-sharing willingness categories

These observations coincide with those obtained by analysing people that live in households with car-sharing subscriptions. These results reinforce the idea that larger households with a reduced number of cars may be making use of these services.

3.2.3.2.8. Main mode of transport

Looking at the main modes of transport used by citizens to go to work or to study, it is observed that those people who regularly travel by bicycle are more inclined to use car-sharing services. Similarly, although to a lesser extent, this is also true for those who usually travel by public transport or on foot. This again confirms the idea of complementarity between car-sharing and active mobility but at the same time some competition with public transport. By contrast, those who usually go to work by car are the most frequent group among those who will not do car-sharing.

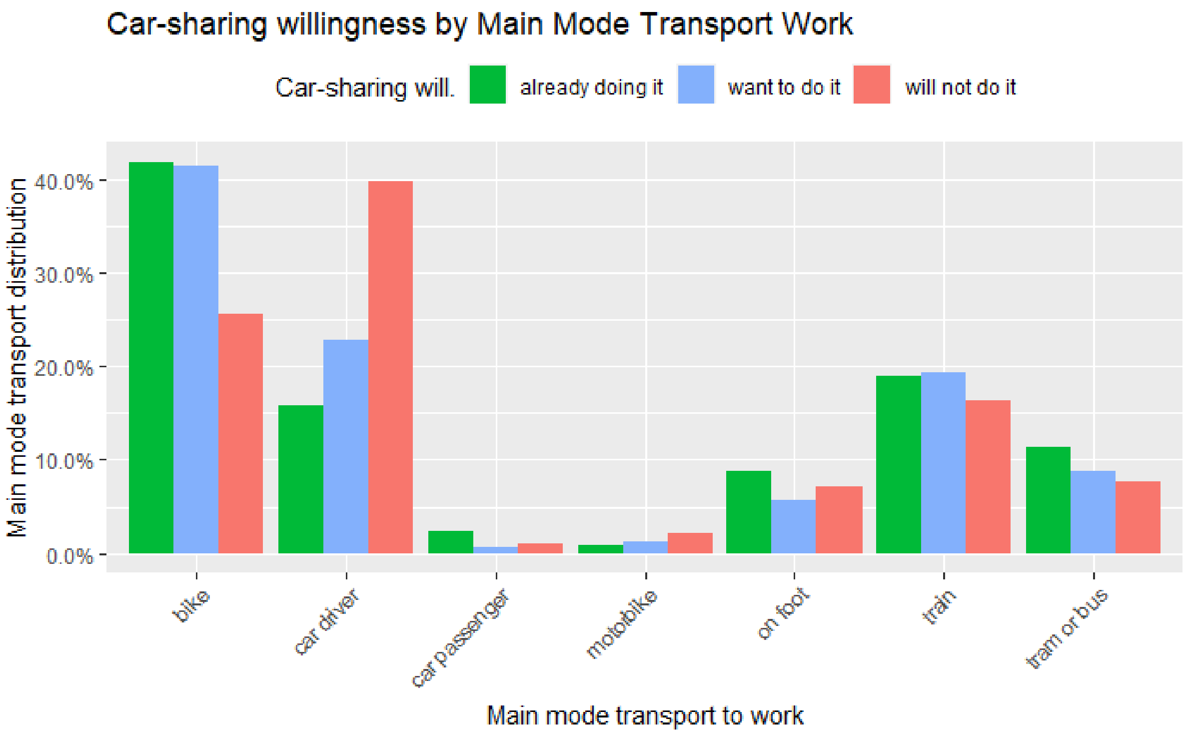


Figure 120. Main mode of transport distribution of car-sharing willingness categories

3.2.3.2.9. City districts

The results of the survey show that willingness to use car-sharing services distribute similarly to the car-sharing subscription: city centre neighbours show a higher interest in using car-sharing. Meanwhile, people living in outer zones, such as Heverlee West and Wilsele Wijgmaal, show no intention to start using this service. This can be in part due to the lower car-sharing supply availability in these districts but also by the composition of the population living in these districts.

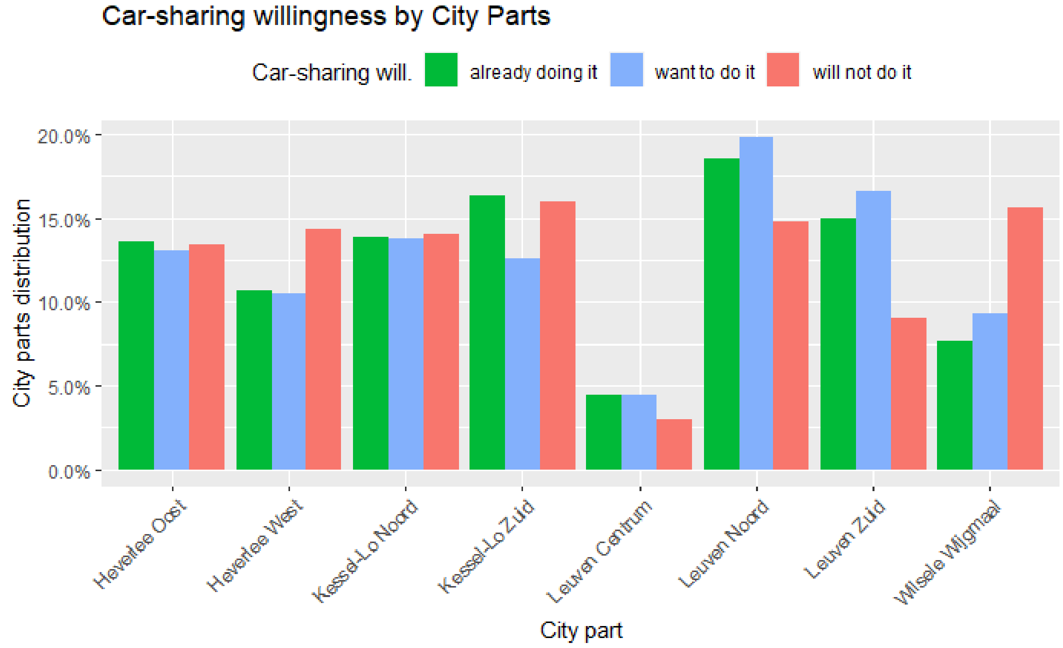


Figure 121. City part distribution of car-sharing willingness categories

3.2.3.2.10. Satisfaction with city centre accessibility by public transport

Firstly, general satisfaction is high and this does not seem to affect a greater or lesser willingness to use car-sharing. However, it can be seen that people with some level of dissatisfaction with the access offered by public transport show a higher proportion of usage of car-sharing. This could suggest that they use car-sharing to compensate for the (perceived) worse accessibility by the public transport system to the city centre. But in any case, this is a weak hypothesis, and further analyses would be required.

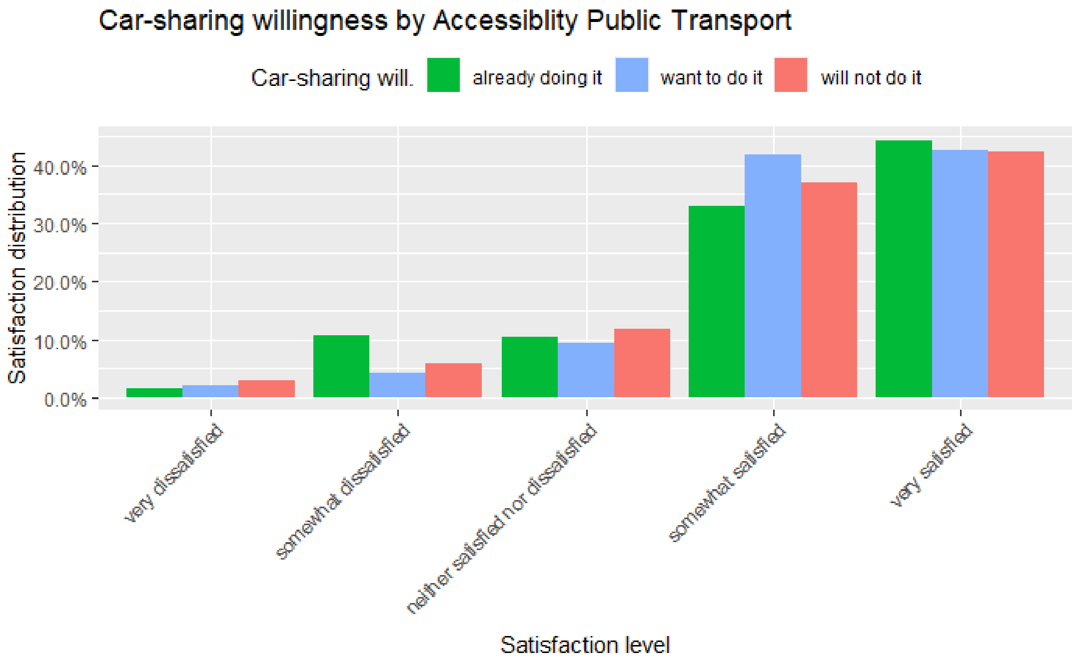


Figure 122. Distribution of satisfaction with the public transport of car-sharing willingness categories

3.2.3.2.11. Vehicle ownership

In a similar way to the analysis of car-sharing subscribers, groups of people that use or want to use car-sharing have on average a lower number of motorised vehicles, both cars and motorbikes. This fact has been clearly observed in other works in the literature (R. Clewlow, 2016) (Becker, Ciari, & Axhausen, 2017).

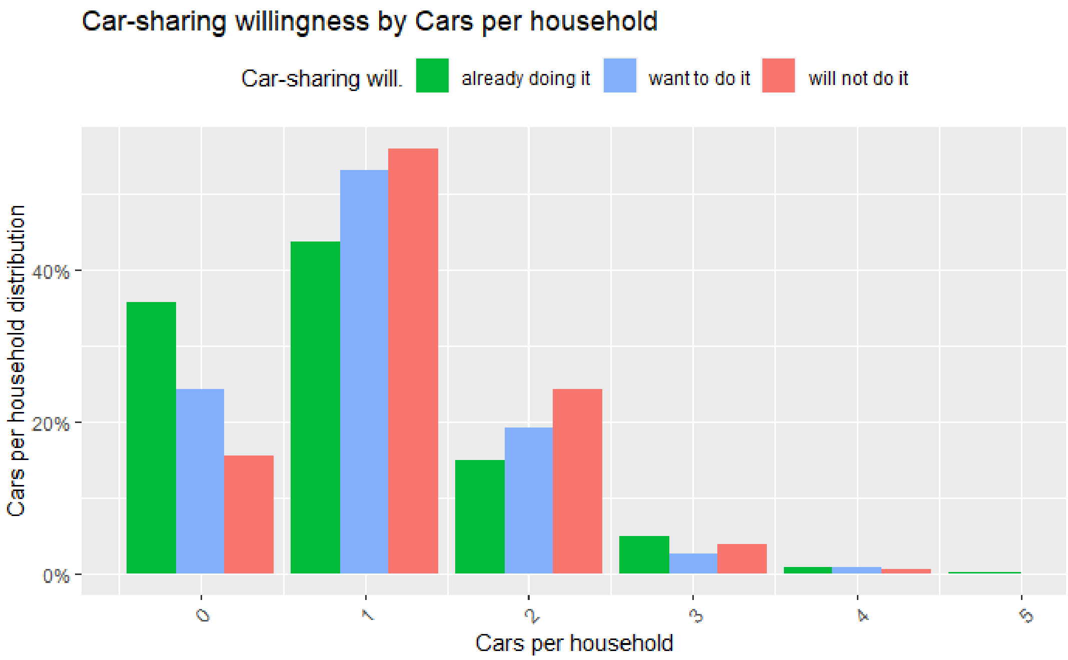


Figure 123. Car-ownership distribution of car-sharing willingness categories

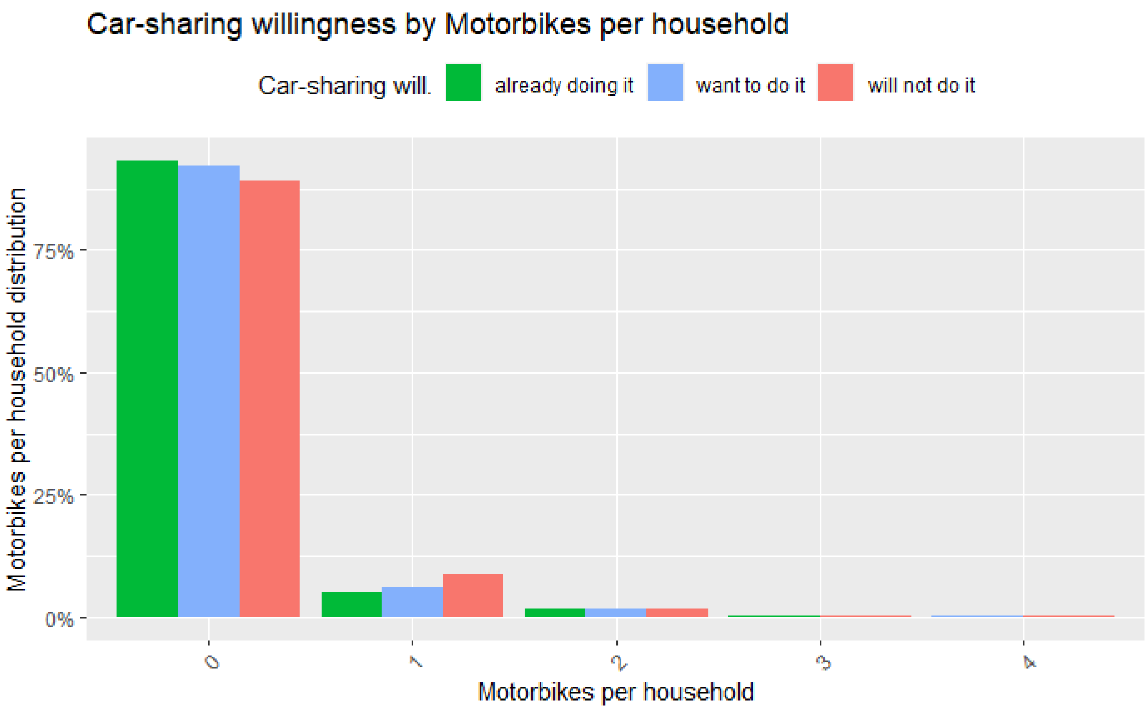


Figure 124. Motorbike-ownership distribution of car-sharing willingness categories

As in car-sharing subscriptions, when we take into account bike-ownership, it is clear that people that already use or want to use car-sharing services tend to own a greater number of bicycles with respect to those who will not.

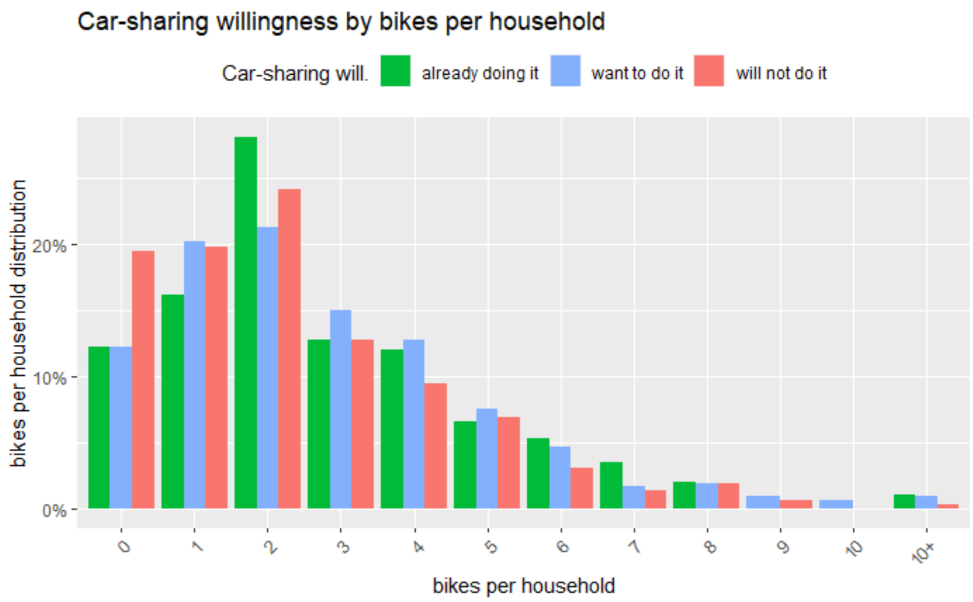


Figure 125. Bike-ownership distribution of car-sharing willingness categories

3.2.3.2.12. Public transport passes

In terms of the number of public transportation passes, the survey results indicate that the group of people who want to use the car-sharing service is the one with the highest average number of public transportation passes. Therefore, this means that many people who use car-sharing does not live in households that regularly use public transport, similar to what happens with users who do not show a willingness to use these services. To conclude this section, it is interesting to compare households with one and two public transport passes, where people who are doing car-sharing and who do not want to do it, are over- and under-represented respectively. The latter fact again underlines the idea that there may be some competition between car-sharing and public transport.

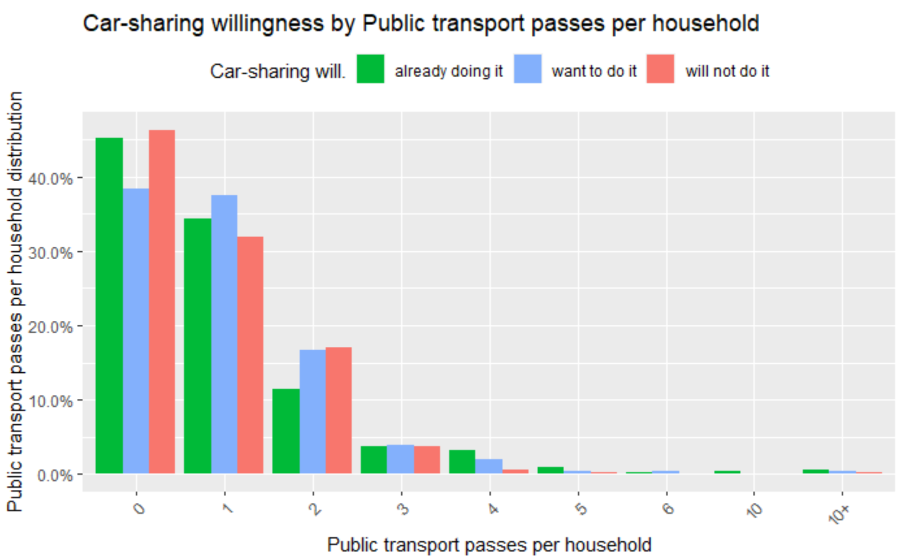


Figure 126. Public transport passes distribution of car-sharing willingness categories

3.2.3.2.13. Car-sharing supply

As with the analysis of car-sharing subscription, we will conclude this sub-section by studying the correlation between the available car-sharing supply and the willingness to use this type of services. More specifically, in this case we have compared the number of car-sharing vehicles available in each city district with the number of people who claim to be doing car-sharing and who want to do car-sharing, respectively.

The following two figures show this correlation. The X axis represents the number of car-sharing vehicles available in each district and the Y axis the number of people already doing car sharing, and that want to do it, respectively. The colour of each bubble also indicates to which district of the city it corresponds, while the size of the bubble is given by the population of that district. In addition, it shows in black the regression line that fits the variables indicated above for the X and Y axes, as well as the R^2 index of this regression.

Similar to what was seen in Section 3.2.2.2.13, the supply seems to be positively correlated with both the number of people who were doing car-sharing and who wanted to do car-sharing, respectively, according to the City Monitor 2017 survey. However, as in the previous case, the correlation is not perfect either, and the supply only explains 22% of the variance in the case of people already doing car-sharing, and 11% in people who want to do car-sharing. In the first of the two cases, we see a relatively similar behaviour to that observed in Section 3.2.2.2.13 with a large difference between Kessel-Lo Zuid and Wilsele Wijgmaal despite having a similar population and supply of car-sharing. In the case of the correlation with people wanting to do car-sharing, the trend is similar, although the differences between the three districts with the largest supply are reduced.

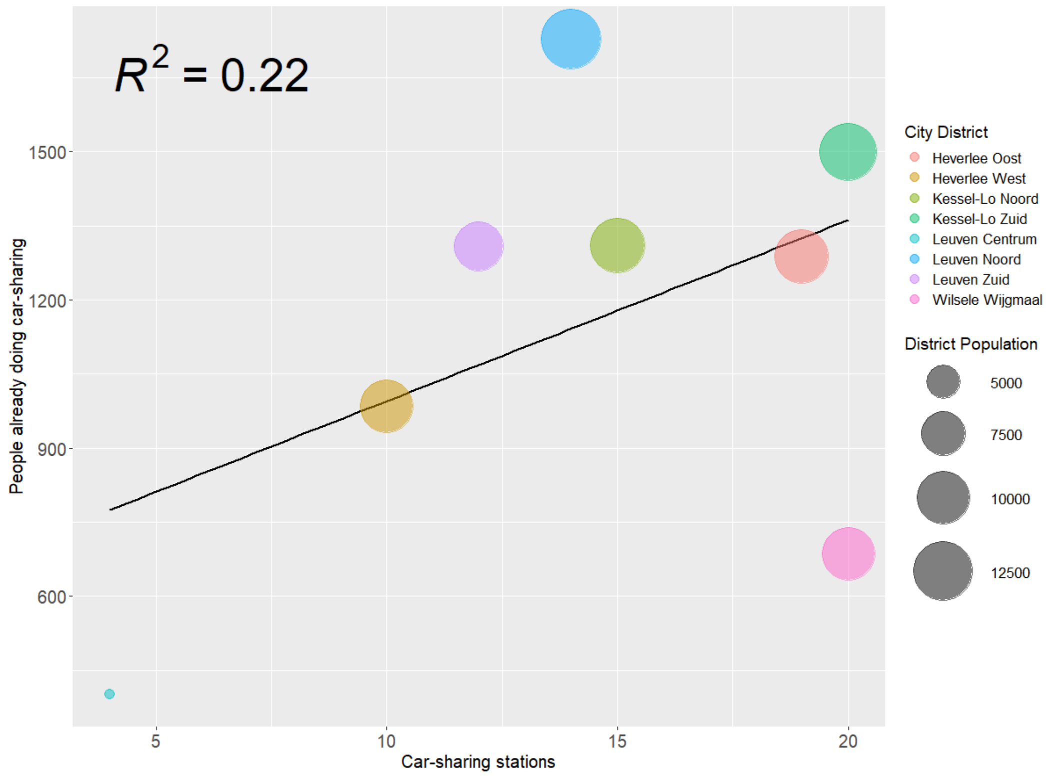


Figure 127 Correlation between car-sharing supply and people already doing car-sharing (District Population)

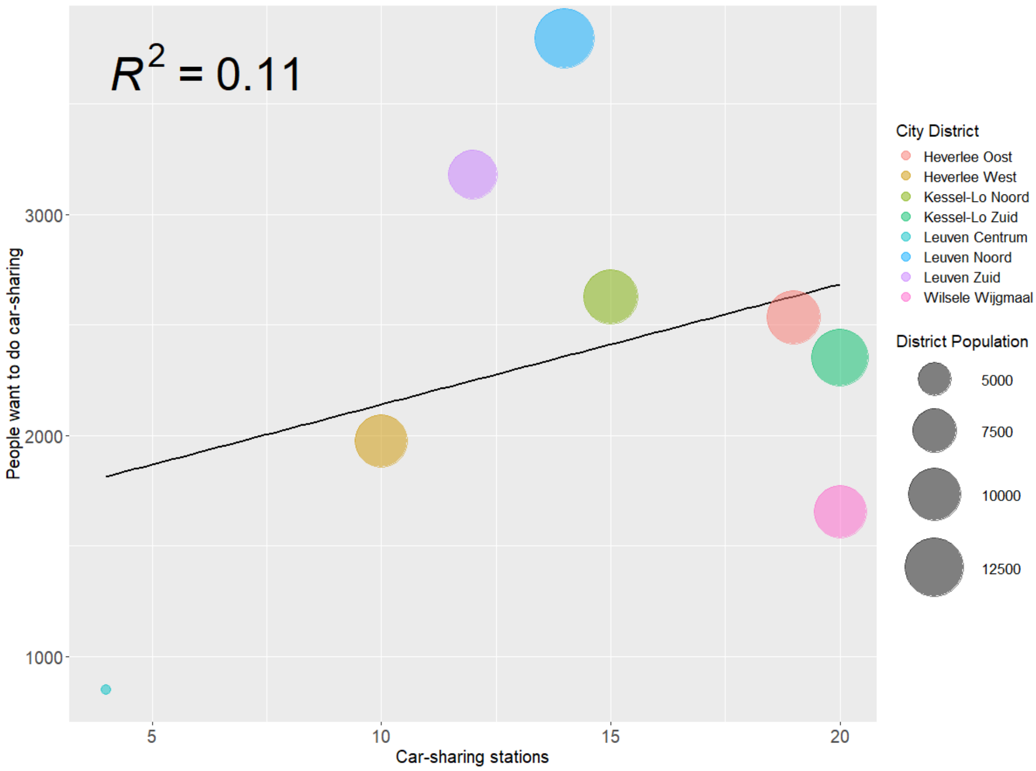


Figure 128 Correlation between car-sharing supply and people that want to do car-sharing (District Population)

To see if other factors can help to better understand the relationship between supply and car-sharing willingness, as we did with the analysis of car-sharing subscriptions, we will introduce a new element in the regression which is the number of people living in households without any private car. Specifically, the adjustment line would now be given by $y=ax+bz+c$, where y would be the number of people who are doing or want to do car-sharing, respectively, z would be the number of people living in households without any private car, while a, b, and c are coefficients. The following two figures show similar bubble charts to the previous ones, with the difference that the size of the bubbles is given by the variable z. The graphs also show the R^2 index for this new fitting. As can be seen, the supply and the number of people living in households without any private car, has a great predictive capacity, already explaining 74% and 86% of the variance of the number of people who are already doing and who want to do car-sharing, respectively.

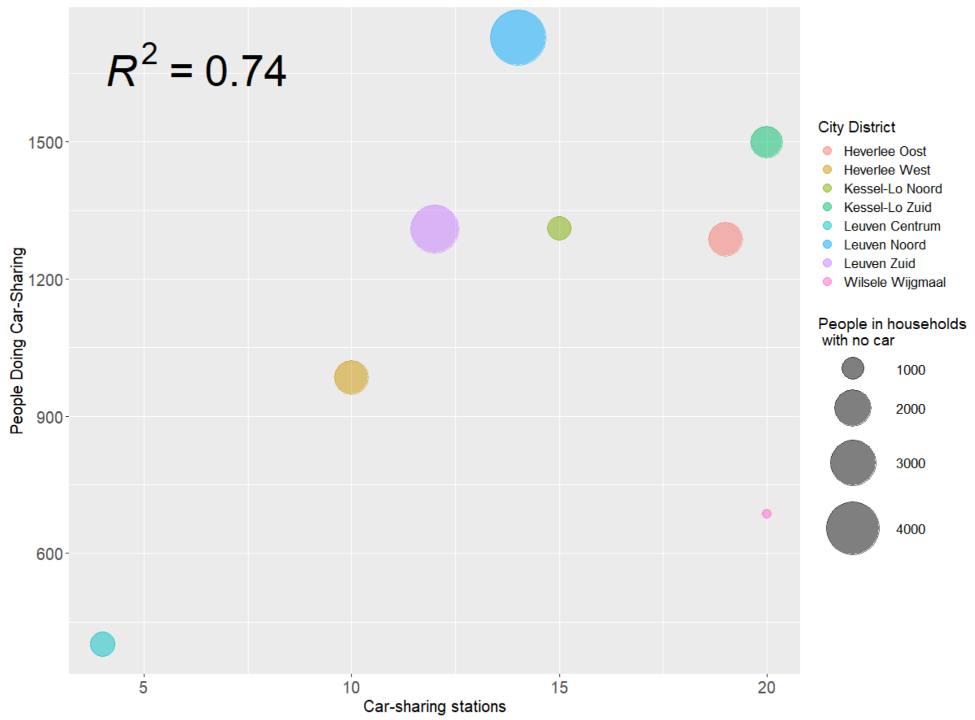


Figure 129 Correlation among car-sharing supply and people living in households with no private car with people already doing car-sharing

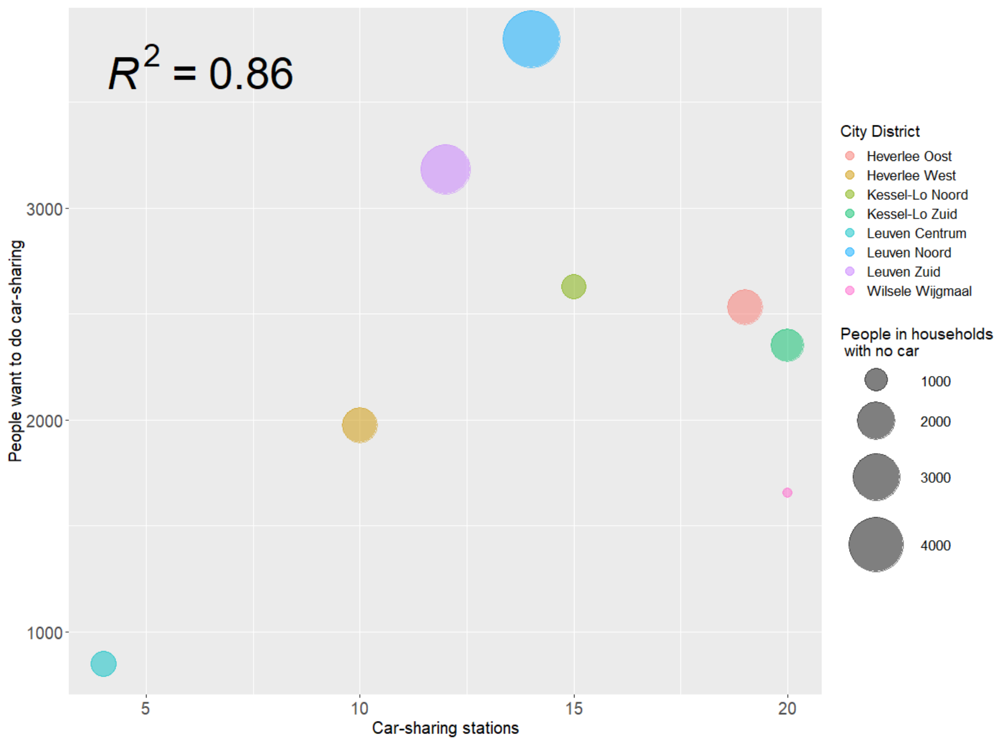


Figure 130 Correlation among car-sharing supply and people living in households with no private car with people who want to do car-sharing

3.2.3.3. Analysis of most relevant factors for car-sharing willingness using white-box Machine Learning

In the previous analysis, we studied the influence that different factors such as sex, education, labour status, etc. have on the probability that people show different levels of willingness to use car-sharing services. Specifically, estimate the probability that they already use it, they want to do it, or they won't do it.

Similar to the study on households with at least one car-sharing subscription, in order to identify which factors have a higher discriminatory power in determining the probability of a person showing a certain level of willingness, we have used the Decision Tree technique. In this case, the dependent variable is a ternary variable that indicates whether the person already uses car-sharing service (AD), wants to do it (WTD) or won't do it (WND). The independent variables are all the factors analysed in the previous section.

Again, it has been necessary to apply techniques to compensate for the unbalance in the amount of data for each class of the dependent variable, since 50,665 responses are available from people who will not use car-sharing services (WND) compared to 19,441 who want to do so (WTD) and 9,479 who already do so (AD). In other words, the three classes are split 11%-23%-66% (AD-WTD-WND). Therefore, by applying the technique called "cost-sensitive", the samples corresponding to WND have been assigned a weight of 1/50,665, while the WTD and AD classes have been assigned 1/19,441 and 1/9,479, respectively.

The construction of the decision tree model has been done in the same way as in the previous section: we have used the Rpart library of R with the default parameters, except for the complexity parameter (cp=0.03) and the parameter that controls the minimum number of observations that must exist in a node in order for a split to be attempted (minsplit=200).

The decision tree obtained can be seen in the following figure. The blue nodes are decision nodes that have a condition associated with them. The blue box indicates the variable involved in the condition, while the white box indicates the condition. The branches that go to the left correspond to the fulfilment of that condition, and the branches to the right the opposite. The orange, grey and green nodes are leaf nodes, which indicate for which of the three possible values of the dependent variable its probability increases with respect to the a priori probability (11%, 23% and 66%, for AD, WTD and WND, respectively). In the white box of the leaf nodes, Prob AD/WTD/WND indicates the probabilities of Already doing it, Want to do it and Won't do it, respectively, while N indicates the number of individuals who meet those conditions according to the survey scaling factors.

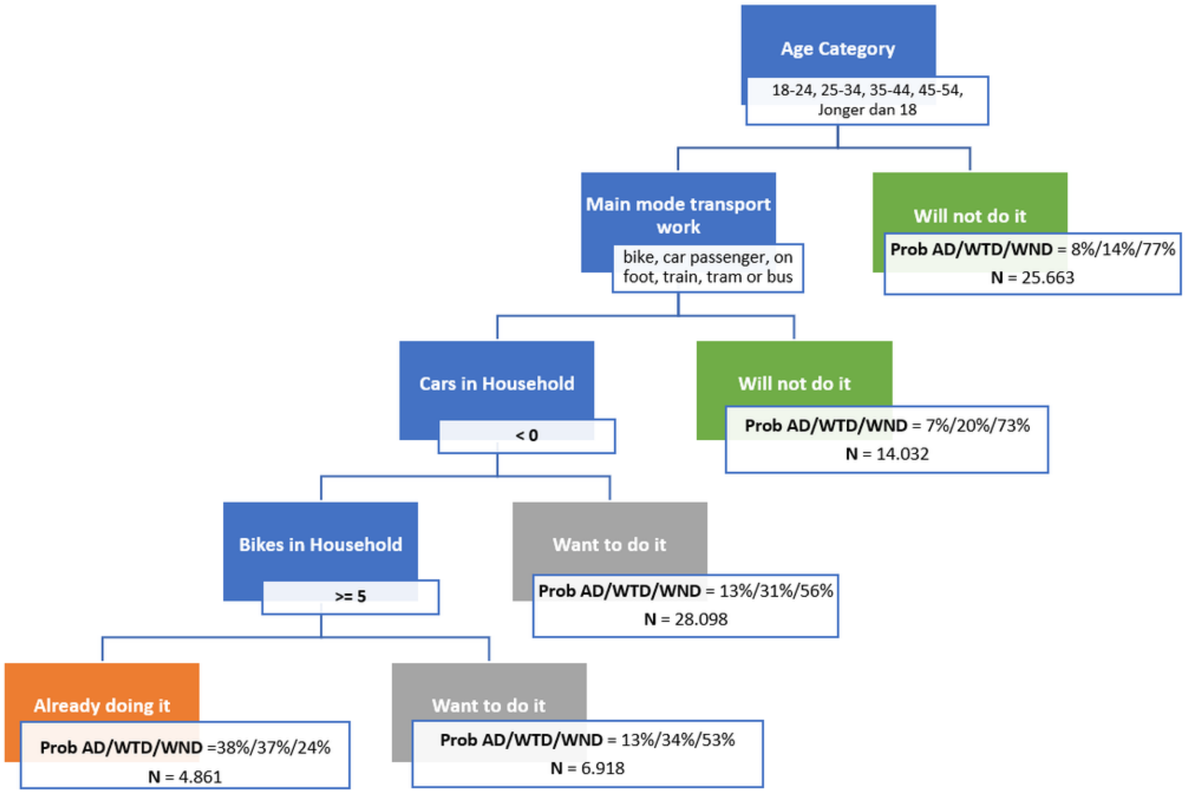


Figure 131. Decision Tree with the most relevant factors for car-sharing willingness levels

Looking at the figure, we see that one of the most relevant variables to discriminate people according to the level of willingness to use car-sharing services is the age. Specifically, people over 55 years old show an increase of more than 10 percentage points in the probability of not considering the use of these services.

The next most relevant variable of those studied is the main transport mode used to get to work. People under age 55 years old who drive cars or motorcycles as their primary mode of transportation are also more likely not to consider car-sharing. However, this increase is somewhat smaller than for the older population group (7 percentage points) and the proportion of people who want to use it also decreases less.

When both of the above conditions are met (population under 55 years old who do not drive motor vehicles as their primary mode of transportation), the probability of showing a willingness to use car-sharing services (WTD) increases by about 50% overall.

In addition, the number and type of cars and bicycles owned by the household make it possible to discriminate between those who want to use this service and those who already do. Specifically, the non-availability of cars in the household, together with having 5 or more bicycles, multiplies by more than 3 the probability of already using car-sharing services.

3.2.4. Analysis of impact of car-sharing in private car ownership

In this section we aim at better understanding the impact that car-sharing may have had in Leuven. To this end, we will make an analysis similar to that presented in Section 3.2.2.3 using Decision Trees. The objective is to analyse which factors have a greater predictive capacity in car-ownership, what effect they have on it, and at the same time to corroborate whether car-sharing is among them. Thus, using again the City Monitor 2017 survey, the dependent variable we have defined has been the number of cars in the household and three possible values have been established: 0 cars in the household, one car in the household and two or more cars in the household. This discretization of the variable has been done in order to avoid excessive unbalance, as households with 3 or more cars are very minority. Thus, the number of individuals in each category is 17,288, 43,280 and 21,272, respectively, which means that their probabilities a priori are 21%, 53% and 26%, in that order. The independent variables considered are the following: sex, age, labour status, education level, household income, household type, household size, motorbikes in the household, bikes in the household, public transport passes in the household, carsharing subscription in the household (yes/no) and carsharing willingness. It is important to note that the variable "City part" has not been included to avoid "masking" the importance of car-sharing, due to the fact, for example, that there is greater availability of these vehicles in certain parts of the city.

The construction of the decision tree model has been done in the same way as in the previous sections, using the Rpart library of R with the default parameters, except for the complexity parameter (cp=0.009) and the parameter that controls the minimum number of observations that must exist in a node in order for a split to be attempted (minsplit=50). Cost-sensitive techniques have been applied.

The decision tree obtained can be seen in the following figure. The colour codes are analogous to previous ones. In this case, in the white box of the leaf nodes, Prob 0/1/2+ indicates the probabilities of having 0, 1 and 2 or more cars in the household, respectively, while N indicates the number of individuals who meet those conditions according to the survey scaling factors.

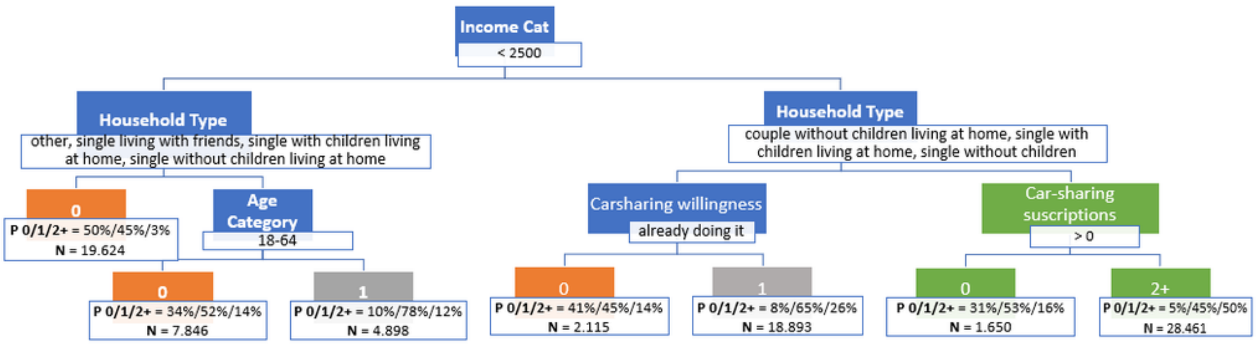


Figure 132. Decision tree for the estimation of car-ownership in Leuven

We can see in the figure that if the income level of the household is lower than 2500€ monthly, then the household type and the age are the best variables to discriminate whether the probability of having 0 or 1 cars in the household increases, respectively. In case the income is higher or equal to 2500€/month, we observe that, after the household type, car-sharing related variables are the one with a more predictive power. In this way, if the household type is couple without children living at home, single with children living at home or single without children living at home, then if the person is already doing carsharing, the probability of having 0 cars in the household is 41%, compared to 9% for those who are not. The probabilities for 1 and 2 or more cars are also lower in this case, indicating a significant trend towards having fewer cars at home from these individuals. However, it is also important to note that only 2116 people fall on this branch of the tree, so their proportion of the total population is small.

In the case of households whose type is couple with children living at home, other, single living with friends or single living with parents or family, something similar to the above occurs, but in this case with the subscription to car-sharing services. If there is at least one subscription of this type in the household, the probability of having no car in the household goes from 5% to 31%, while the probability of having two or more cars goes from 50% to 16%. But as in the previous case, the proportion of people in households with a car-sharing subscription is small, so the results should be taken carefully.

These results confirm the relationship between car-ownership and car-sharing, although with the available data it is not possible to establish the direction of causality. Specifically, if people who do car-sharing reduce their private car ownership, or if, on the contrary, people who do not have a car make greater use of car-sharing services, which can be understood in some way as a certain incentive towards car use. This point will be discussed in more detail in the Section 4.2.

3.3. Regensburg Case Study

3.3.1. Data sources

The main data sources used in Regensburg case study are shown in Table 35. A more detailed description of the data sources can be found in deliverable D3.1.

Data source	Data inventory code	Comments
Mobility household survey	TD_006	Regional mobility household survey conducted in 2018.
Station-based car-sharing information from operation data	TS_004 and TD_009	Trip information obtained from operational data of two car-sharing companies in the Regensburg metropolitan area: one provides a city-level services and the other one regional-level. The data used for the analysis range from November 2016 to November 2019.

Table 35. Main data sources used in Regensburg case study

In the Regensburg Case Study, the mobility survey available corresponds to the study "Mobility in cities – SrV" (TD_006) performed by TU Dresden in 2018, whose results were available for Regensburg at the end of 2019. This is the eleventh edition in a series of studies that began in 1972 that uses a household survey to obtain data on travel behaviour in selected cities and regions of the Federal Republic of Germany with Regensburg being one among them. For Regensburg, this survey took place from February 2018 until January 2019 on typical working days (Tuesday, Wednesday and Thursday), and only households in the Regensburg urban area were surveyed. A total of 2,501 people in 1,116 households were successfully interviewed. For this study, the Regensburg urban area was divided into five sub-areas: centre, north, south, east and west. The survey contained questions related to emerging mobility solutions, including the general use of car-sharing. Therefore, the analysis of the results of this survey allows obtaining a profile of the socio-demographic characteristics of the users of these services in relation to their frequency of use of the car-sharing service.

In the Regensburg metropolitan area there are two public station-based electric car-sharing systems: one called “das Stadtwerk.Earl”, that is deployed in the city of Regensburg and run by the company das Stadtwerk Regensburg Mobilität GmbH (a wholly-owned subsidiary of the city of Regensburg); and another one called “KERL”, deployed in the district of Regensburg and run by the company Kommunale Energie Regensburger Land eG. As shown in Figure 133, these are two small services. The first one has eight stations distributed in the city of Regensburg (blue markers), whereas the second one has ten stations distributed in different cities from the District of Regensburg (green markers). From now on, we will refer to them as “City” and “District” services, respectively.

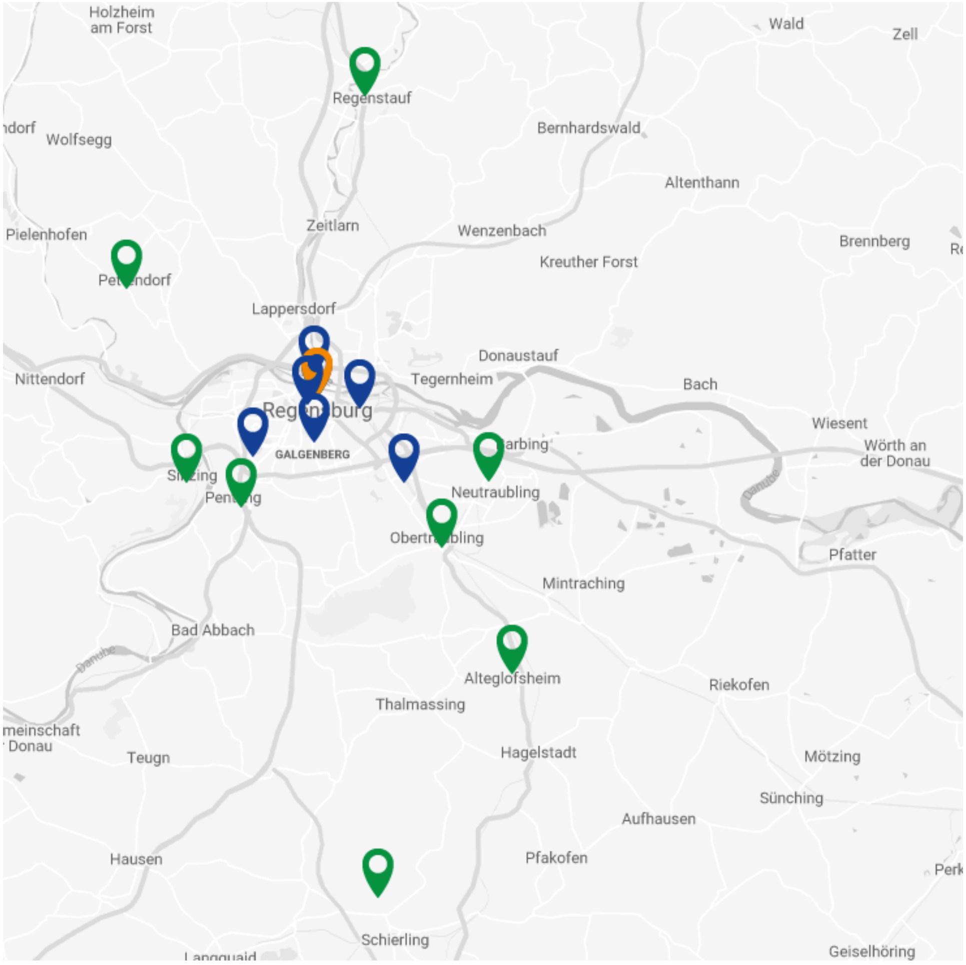


Figure 133. Car-sharing stations in the area of Regensburg (blue for the city service and green for the district service)

For the Regensburg Case Study, supply and demand data from both the car-sharing services are available (TS_004 and TD_009, respectively). The temporal scope of the dataset ranges from November 2016 to November 2019, containing 19,355 trips. The data were extracted from the operating tool used for providing the customer service and for tracking the vehicles, and it includes the following information for each trip: start date, start time, return date, return time, used time, pick up location, return location, make, model, registration, unit, trip length, estimated trip length, start booking (km), return (km), type, status, booking edited, booked start date, booked start time, booked return date, booked return time and booked trip. After selecting only trips made by users (e.g., discarding maintenance trips) and removing those with no distance travelled or no duration, the resulting dataset had 10,754 trips, of which 8,352 belong to the City service and 2,402 to the District service.

3.3.2. Characterization of car-sharing usage frequency

3.3.2.1. Applied methodologies

The characterization of car-sharing users is based on an analysis of the answers obtained from the "2018 Mobility in cities – SrV" survey for the question "In the past 12 months, how often have you used a Car-Sharing vehicle (as driver or passenger) in general?". The possible answers were:

- Daily or almost daily
- 3 to 4 days per week
- 1 to 2 days per week
- 1 to 3 days per month
- On 1 or 2 days per quarter
- Rare
- Never

According to these answers, the surveyed individuals are categorized according to the frequency of use that they have reported, and are compared and analysed under different socio-demographic, economic and transport habit parameters, with the aim of profiling the characteristics that define these types of users. The number of samples after weighting in each category are the following: Daily or almost daily (N=0), 3 to 4 days per week (N=1), 1 to 2 days per week (N=15.3), 1 to 3 days per month (N=39.3), On 1 or 2 days per quarter (N=34.4), Rare (N=81), Never (N=2,327.8). Given that the categories “Daily or almost daily” and “3 to 4 days per week” have 0 and 1 samples, respectively, we decided to discard them from the analysis to facilitate the understanding of the results.

3.3.2.2. Result analysis

3.3.2.2.1. Sex

With respect to the sex of the users, firstly it can be observed that women are over-represented within the group of people who never use car-sharing. Or to put it another way, men use the service more.

The most frequent user group is gender-balanced. However, in other groups, men are the majority. Therefore, the survey indicates that, in general, in Regensburg men tend to use more car-sharing services. Unlike the one in Leuven, these results are in line with literature in this field (Becker, Ciari, & Axhausen, 2017; Yoon, Cherry, & Jones, 2017).

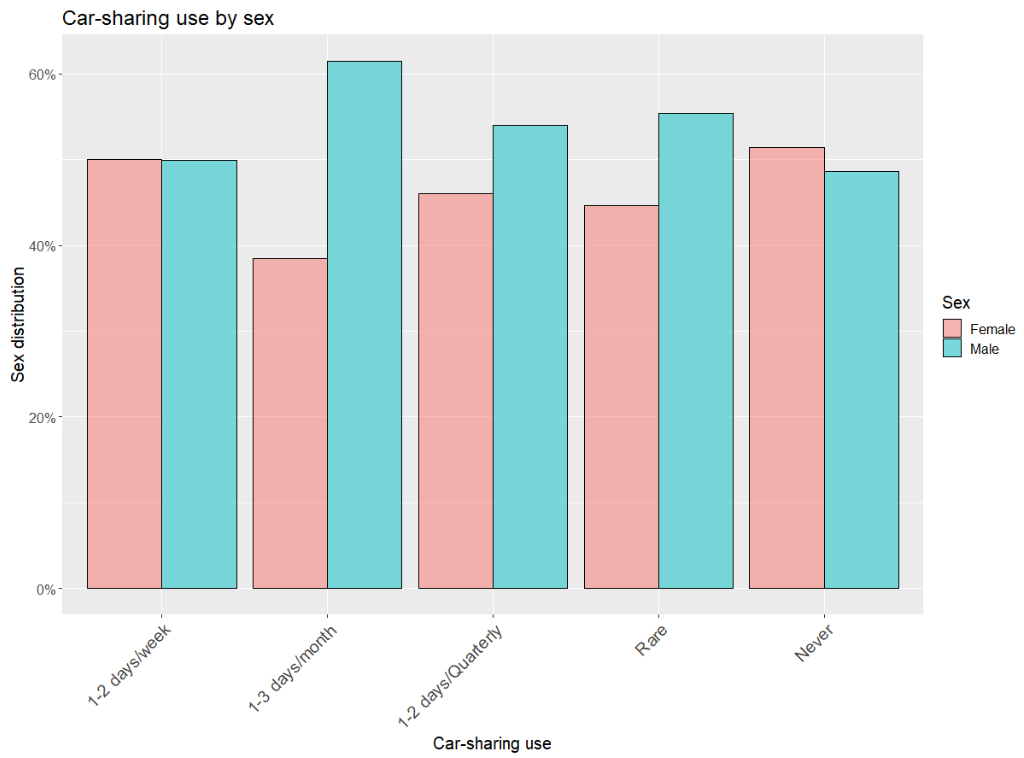


Figure 134. Sex distribution of car-sharing usage frequencies

3.3.2.2.2. Age

The age is again an important factor influencing the frequency of use of car-sharing. Several facts can be observed:

- Firstly, users in the 15-24 age range are over-represented in the rest of the use patterns, especially in those with the highest frequency of use.
- On the other hand, the group of 25 to 44 years old appears as the second group in the most frequent patterns and leads the categories of occasional use.
- And finally, from age 44 onwards, car-sharing use drops significantly, especially from age 65 onwards where it is almost non-existent.

These data indicate that the main users are young people, in the range 15-44, although, according to the results, one can differentiate between a younger group (15-24) who use car-sharing more frequently and an older group (25-44) who use it more occasionally. These results are compatible with those reported in several studies, such as those conducted in the area of San Francisco (United States) (R. Clewlow, 2016) or Basel (Switzerland) (Becker, Ciari, & Axhausen, 2017), indicating that, currently, the attractiveness of car-sharing decreases significantly as the age of the individual increases.

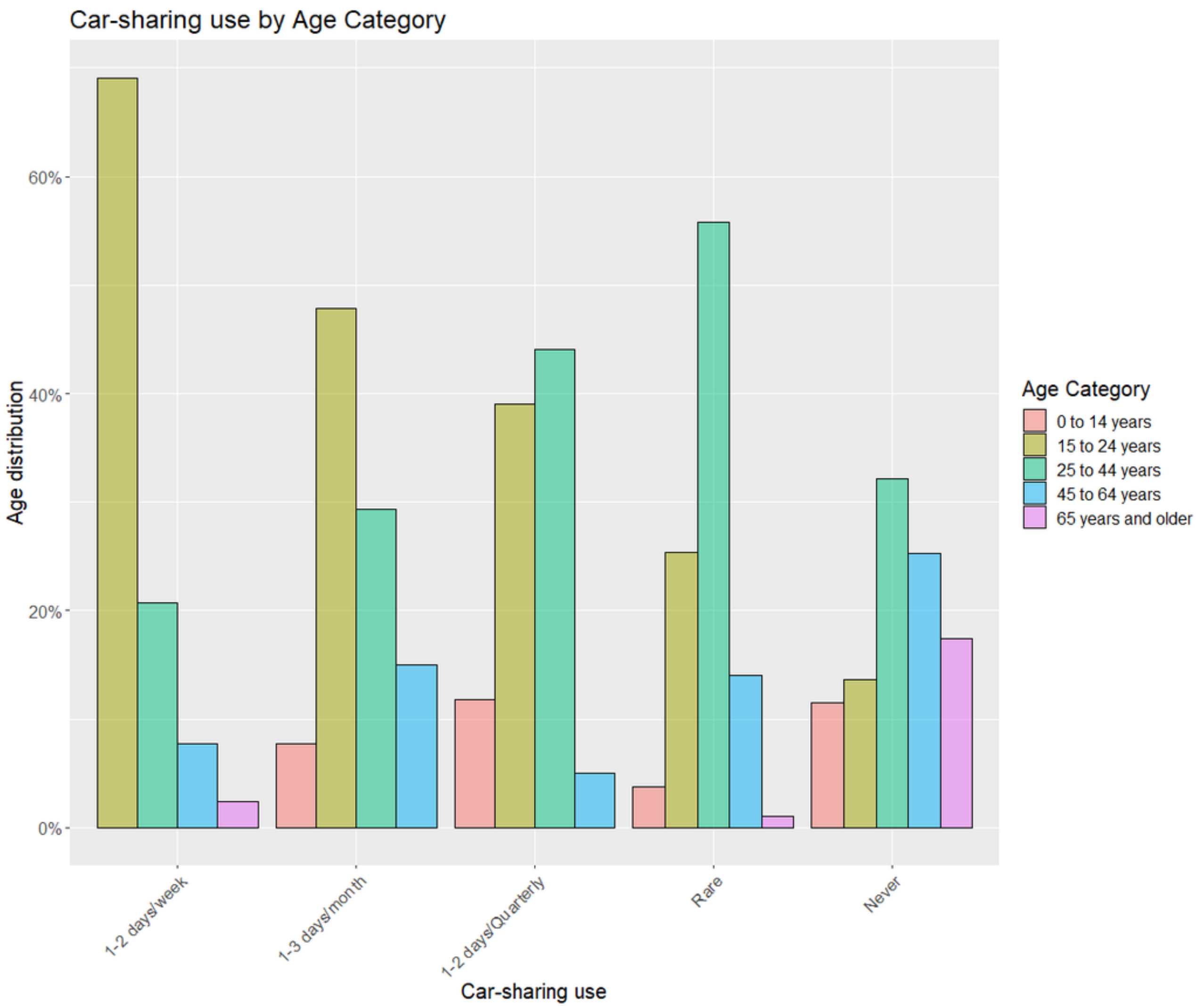


Figure 135. Age distribution of car-sharing usage frequencies

3.3.2.2.3. Labour status

Taking as a control group the category of users who never use car-sharing (also very similar to the general distribution of the population according to their employment situation, because of its big size), the main insights that we can extract are the following:

- The first thing the survey data indicates is that in the most frequent car-sharing user group, students and full-time workers are clearly over-represented. In fact, students are over-represented in all categories that entail some frequency of car-sharing use. These results are in line with what we saw above with respect to the user's age.
- It is also interesting to note how part-time workers tend to use car-sharing but mainly with an intermediate frequency.
- Again, survey data confirm that the older population, mainly retirees, barely make use of car-sharing.

These results are consistent with those that can be found in the literature, such as the study conducted in Basel, Switzerland, which also observed a special over-representation of students and workers, especially self-employed workers (Becker, Ciari, & Axhausen, 2017).

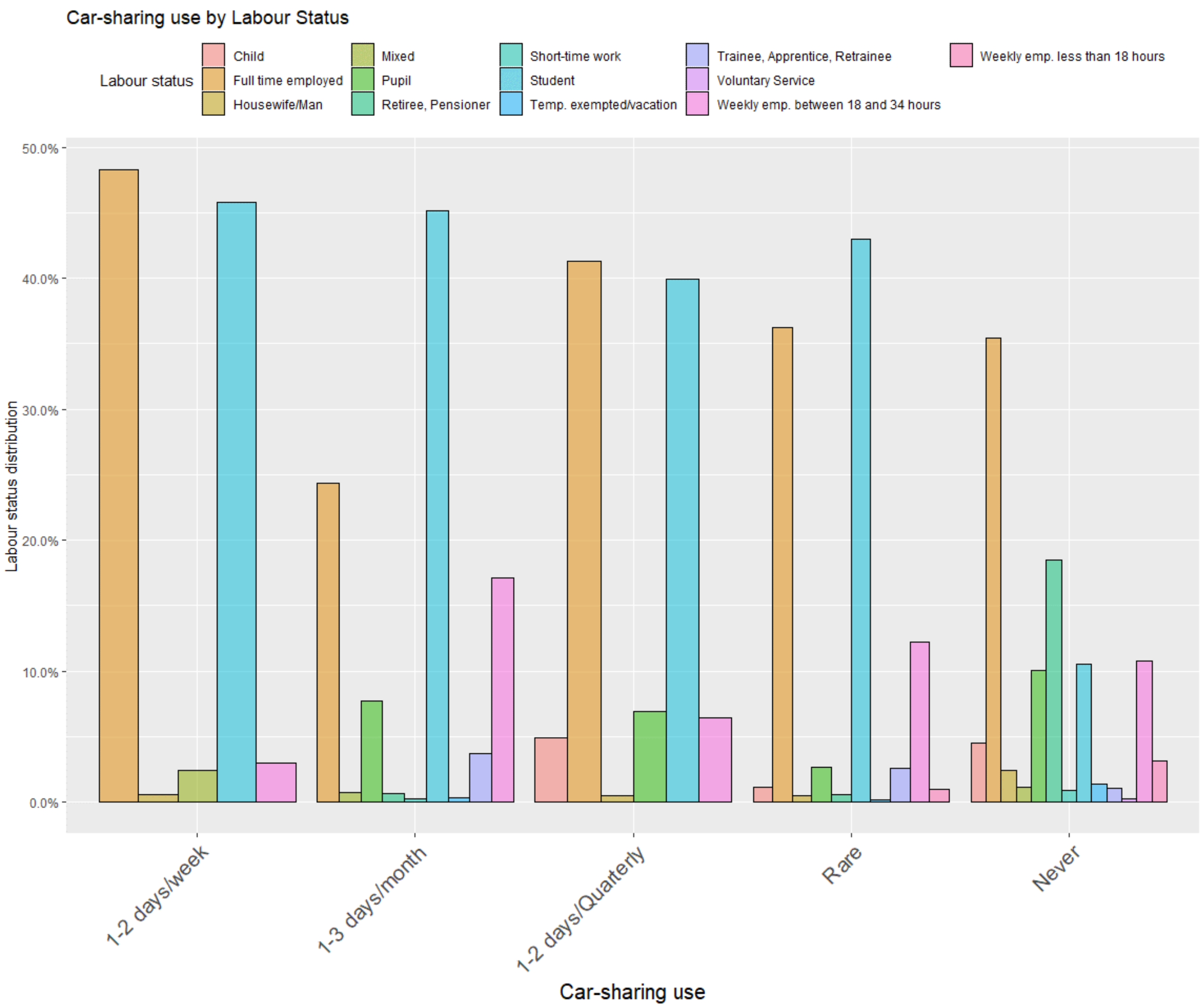


Figure 136. Labour status distribution of car-sharing usage frequencies

3.3.2.2.4. Education

When analysing the different car-sharing frequency usage patterns according to the level of studies of the people who are part of them, it can be seen that having a university degree is a general characteristic of almost all car-sharing users. This observation is entirely consistent with that which can be found in the scientific literature, where, in general, more than 70% of car-sharing service subscribers have a Bachelor's degree or higher (R. Clewlow, 2016) (Becker, Ciari, & Axhausen, 2017) (Yoon, Cherry, & Jones, 2017).

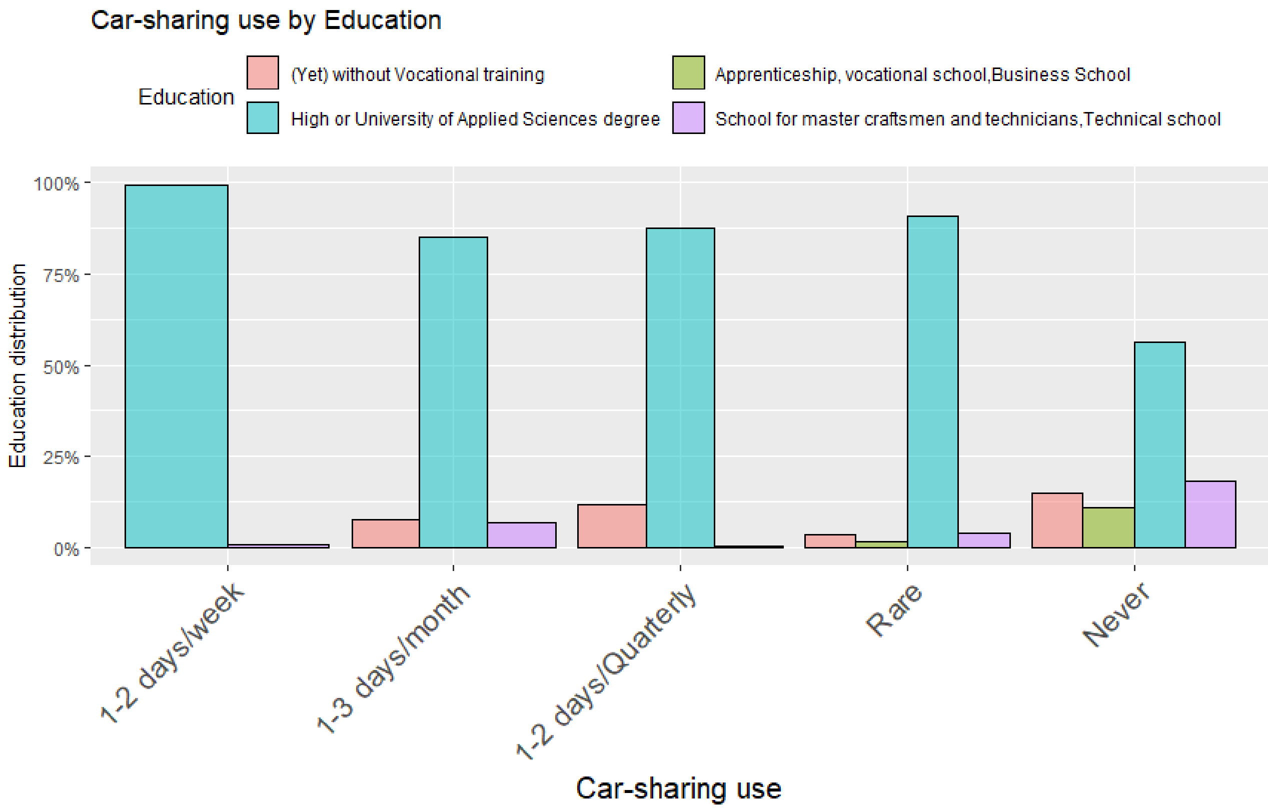


Figure 137. Education level distribution of car-sharing usage frequencies

3.3.2.2.5. Incomes

When analysing the different frequencies of car-sharing use according to the incomes of the users, it can be seen that among the most frequent users there is a significant proportion of people with very low incomes. This may be, again, linked to the age variable and corresponding to students. In this group of frequent users, there are also intermediate salaries that may correspond to young workers.

On the other hand, it is observed that, in the more sporadic use patterns (1-2 days/Quarterly and Rare), there is a great proportion of people with high salaries.

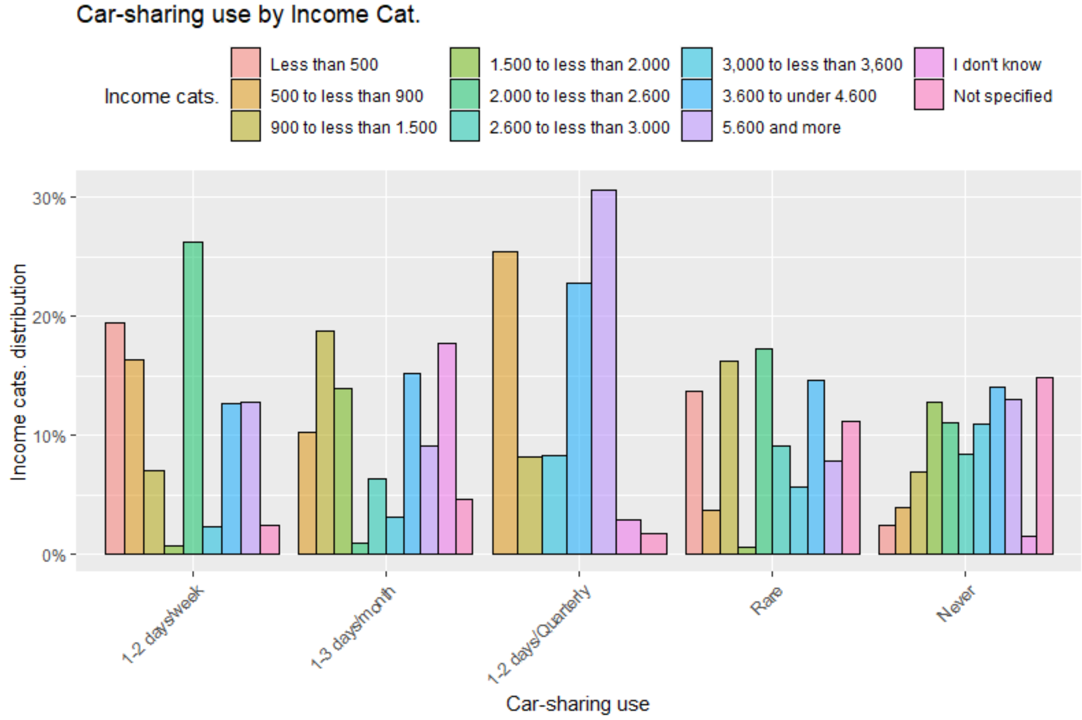


Figure 138. Incomes distribution of car-sharing usage frequencies

3.3.2.2.6. Household size and type

Analysing the different types of car-sharing users according to the size of the household in which they live, it can be seen that households with only one inhabitant are over-represented (with respect to their proportion in the total population) in the most frequent use patterns. And, similarly, the proportion of households with 2 inhabitants also stands out in the pattern of sporadic use.

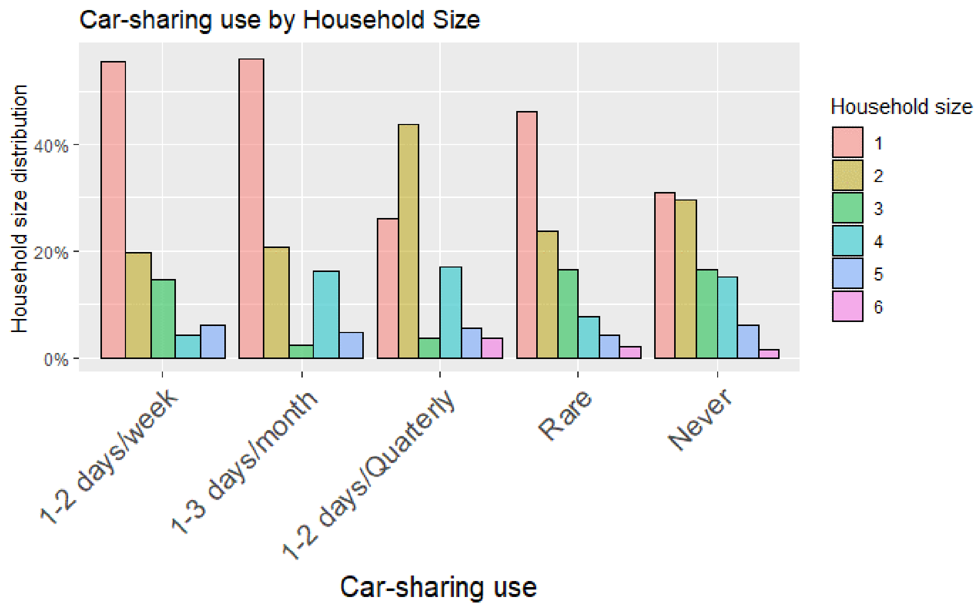


Figure 139. Household size distribution of car-sharing usage frequencies

Doing the same analysis according to the type of household, it can be suggested that the first case would seem to correspond to people living alone under 65 years of age (mainly singles), while the second case would seem to be linked to couples without children. It is also observed that families with children tend not to use car-sharing services much. Finally, it is again observed that older people (singles over 65 or multi-person without children) tend to never use car-sharing.

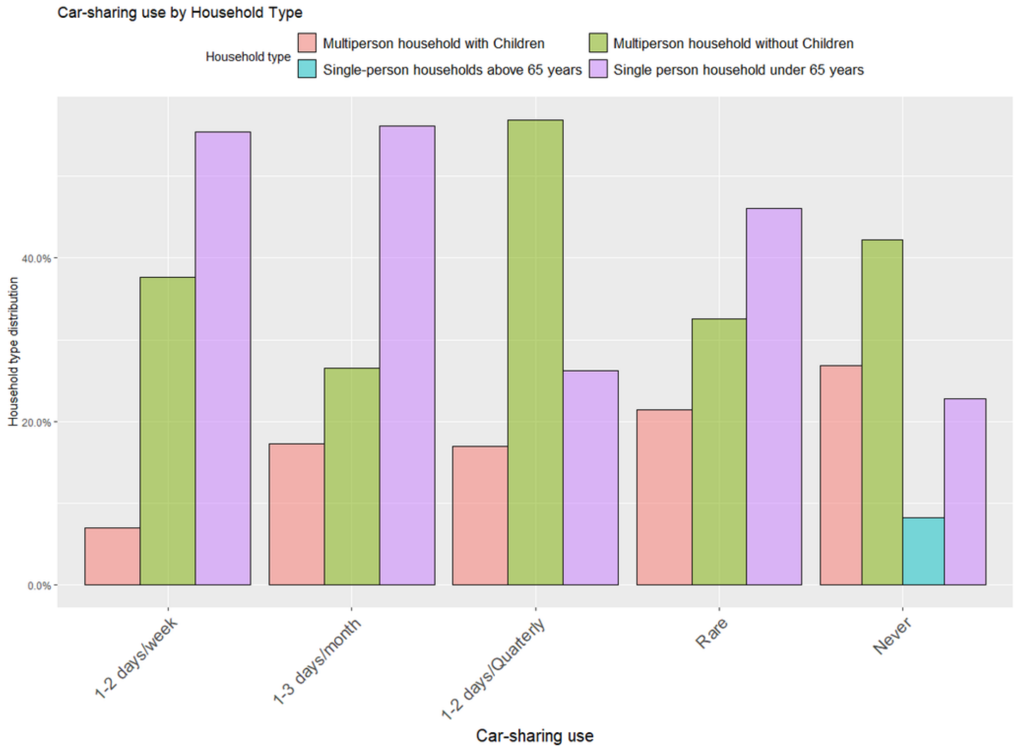


Figure 140. Household type distribution of car-sharing usage frequencies

3.3.2.2.7. Vehicle ownership

By analysing the frequency of use of car-sharing services in relation to the types and quantity of vehicles owned in the household, interesting observations are obtained.

Firstly, the lack of car ownership is strongly correlated to the use of car-sharing, as shown by the over-representation of this group in all usage patterns. In contrast, the population that has some car in the household tends to be under-represented in the groups that entail some use of car-sharing. The only exception to this is the category of three vehicles for “1-2 days/week”, but in this case, it may be due to an artefact of the survey since the number of samples in this category is low.

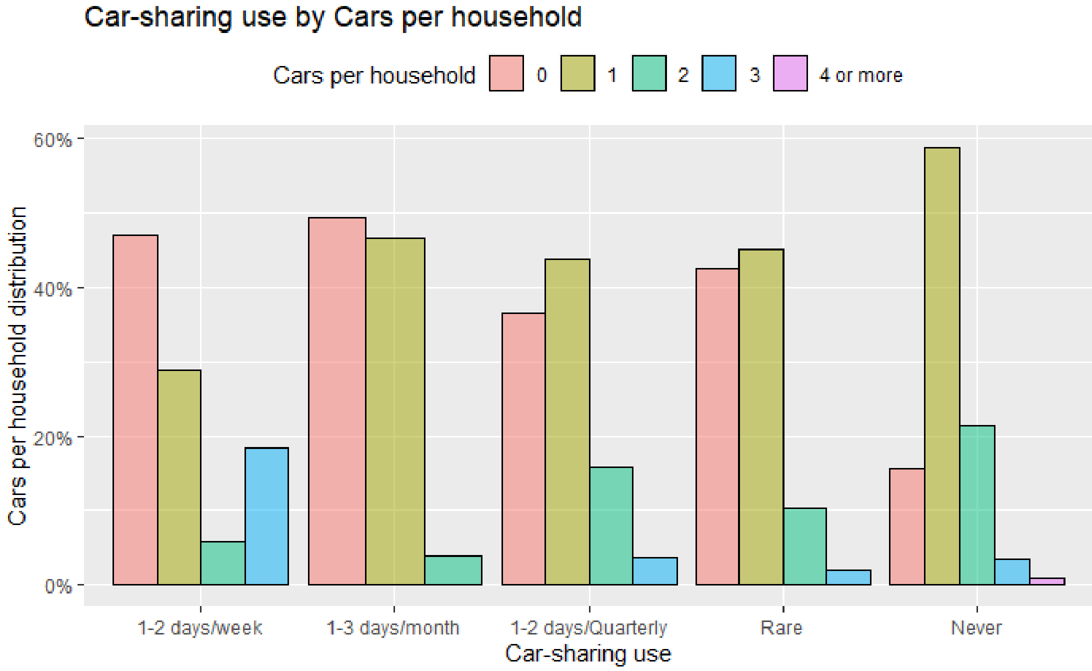


Figure 141. Car-ownership distribution of car-sharing usage frequencies

With respect to bicycles, taking the category Never as a control group, we see that the group with no bike in the household is under-represented in all the categories that imply some use of car-sharing. On the contrary, households with one bike are over-represented in the categories we have just mentioned. This may be linked to the fact that most households using car-sharing are one-person homes. For a larger number of bicycles in the household, there is no clear trend, although most of them are over-represented in the case of infrequent use of car-sharing ("Rare" and “1-2 Days/Quarterly”).

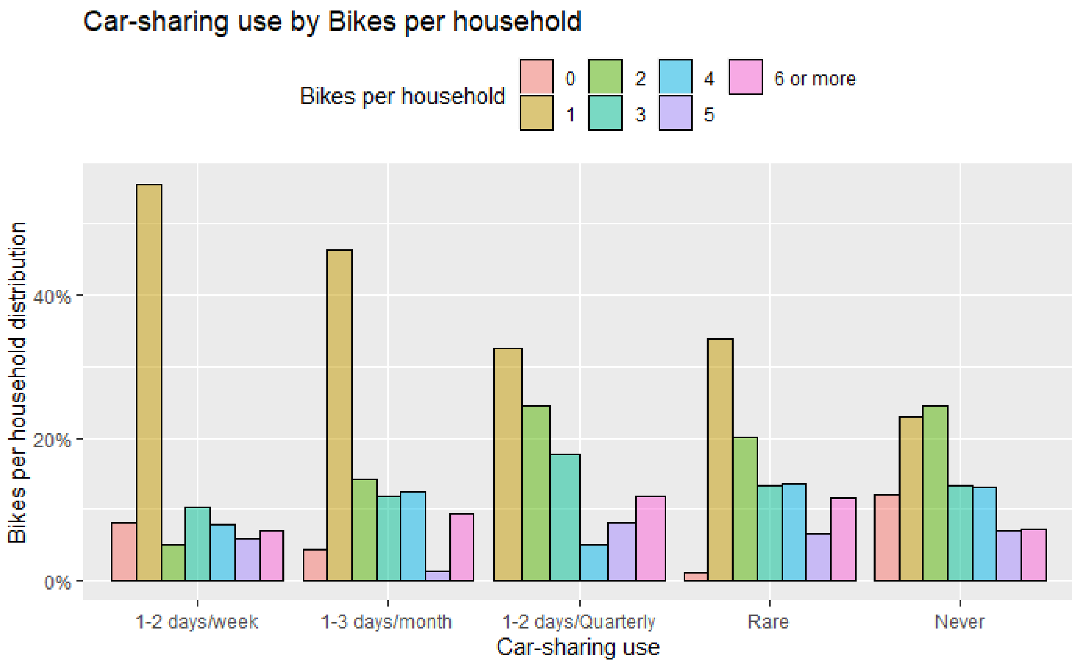


Figure 142. Bike-ownership distribution of car-sharing usage frequencies

3.3.2.2.8. Public transport passes

From the survey data, possible relationships between possession of different types of public transport passes and car-sharing patterns can be analysed. Taking again the category “Never” as the control group, the main results to highlight is the over-representation of individuals with “Job/Semester ticket” being most of them students, which is in line with previous results. If we focus on people who use "Single ticket", which suggests infrequent use of public transport, we see that they are over-represented in the categories "1-3 days/month" and "1-2 days/Quarterly". This seems to indicate that 30% of people who use car-sharing services between one and three times a month and one or two days per quarter, make rather sporadic use of public transport. Finally, the other interesting ticket to be analysed is the multi-trip. In this case, we see that except in the most frequently used category (where, as we have already mentioned, there is a certain risk of artefacts in the survey), in the rest it is under-represented. This may suggest that the car-sharing system may be affecting people who make intermediate use of public transport, although this fact cannot be confirmed with current data.

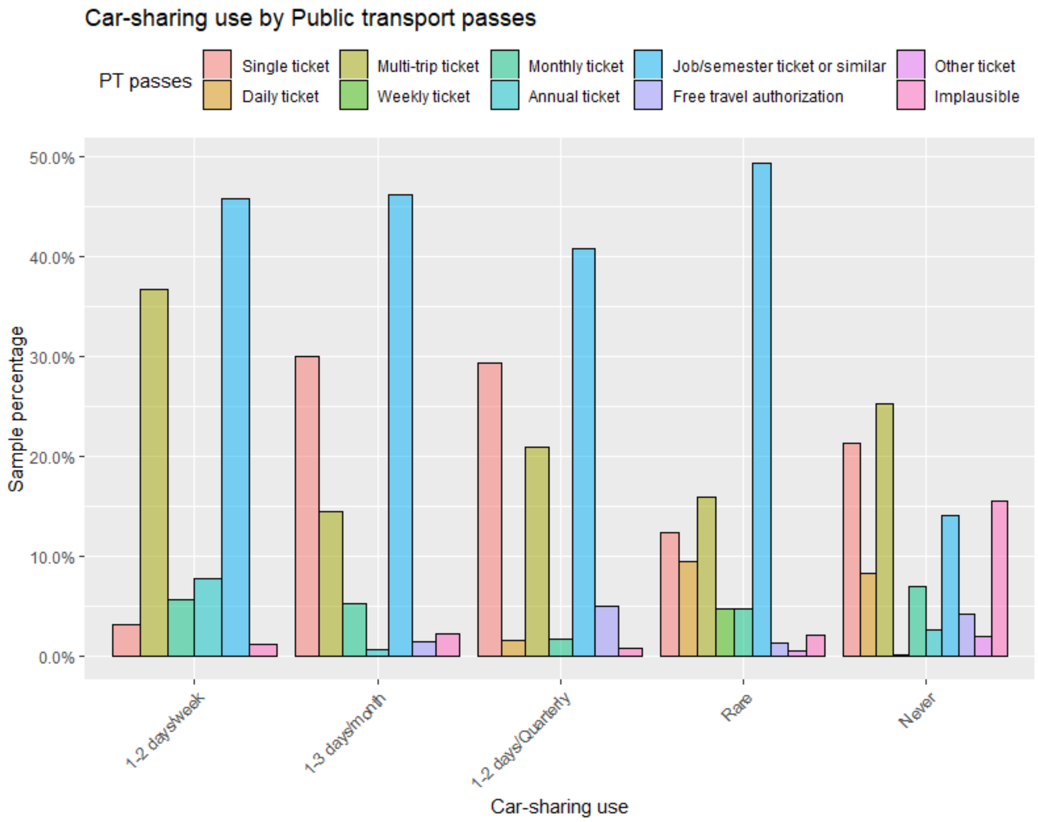


Figure 143. Public transport passes distribution of car-sharing usage frequencies

3.3.2.2.9. Modal share

As mentioned above, the household survey available in Regensburg also provides information about the trips performed on Tuesday, Wednesday and Thursday of surveyed individuals. The survey collects data on trips such as origin, destination, departure time, arrival time, distance travelled, purpose or the main transport mode used, among others. Using this last data, we have estimated the modal-share for the different categories of car-sharing use frequencies that we have been analysing in this section. The results obtained are shown in the figure below. In it, we can see how in the “Never” category the most frequent modes of transport for Regensburg are the car as a driver, on foot and by bicycle, in that order, and in a smaller scale the bus and the private car as a passenger. If we take this category as a control group, we see that the modal-share of car-sharing users is different, although this difference is not homogeneous but varies with the frequency of use. In general, we see that bicycle use is more frequent in all cases, except for the category "1-2 days Quarterly". However, the use of the private car behaves inversely. Specifically, the share is lower in the two categories of most frequent car-sharing and in "Rare", and while similar to the control group for those who travel on this service 1 or 2 days Quarterly. Another interesting aspect observed is the reduction in the share of trips on foot almost inversely to the increase in the frequency of car-sharing use. This is probably due to greater use of bicycles and buses. In fact, the latter shows a completely opposite behaviour. In other words, its share increases with the frequency of use of car-sharing services.

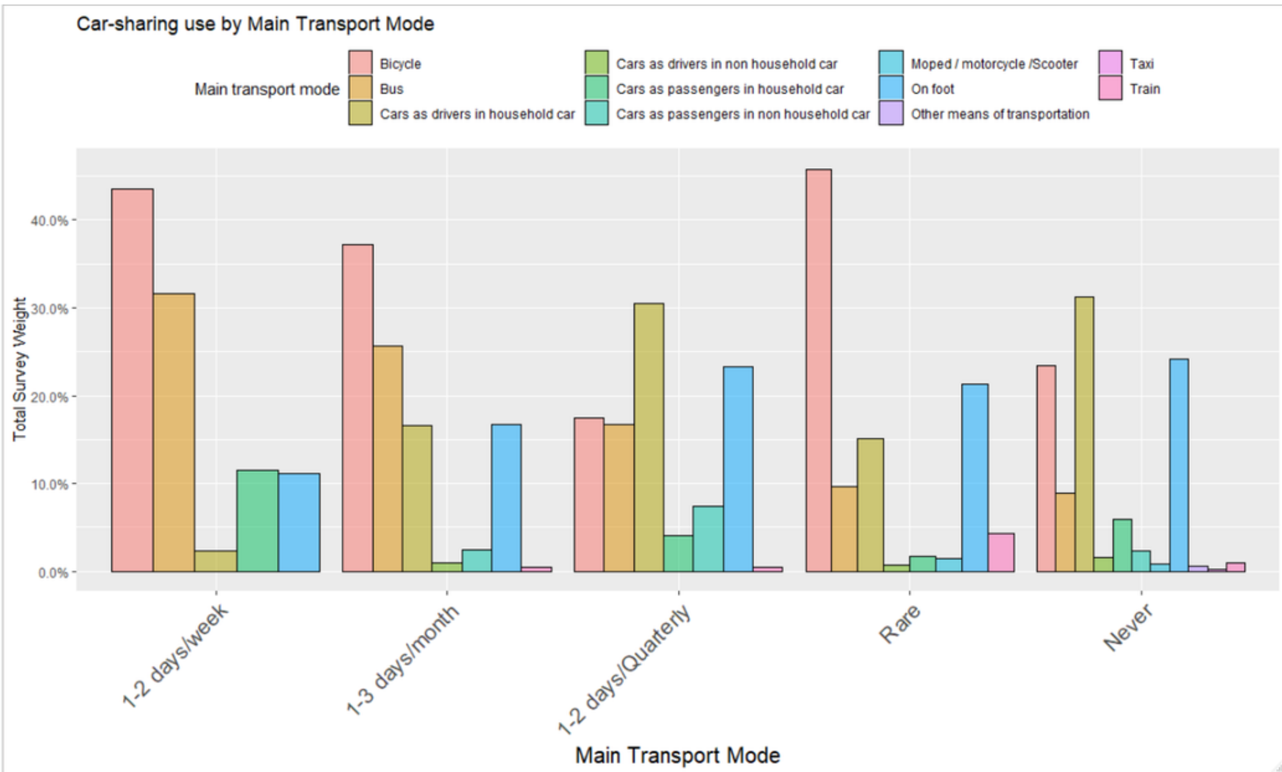


Figure 144. Modal-share by car-sharing usage frequency

3.3.2.3. Analysis of most relevant factors for car-sharing adoption using white-box Machine Learning

As in the Leuven Case study, we will make use of Decision Trees in order to establish the factors that have higher discriminatory power for determining the adoption of the car-sharing services in Regensburg and also their frequency of use.

In this case, the dependent variable, related to the frequency of use of these services as we have just mentioned, has three possible categories: "Do not use", which corresponds to the category "Never" in the analyses carried out in the previous section; "Low frequency", which groups together the categories "Rare" and "1-2 days/Quarterly"; and finally "Medium-High Frequency", which corresponds to the two categories of most frequent use, "1-3 days/Month" and "1-2 days Week".

It has also been necessary to apply "cost-sensitive" techniques in the decision trees to deal with the high class-imbalance, since the number of samples after survey weighting is 55.6, 115 and 2,328, for Frequent, Infrequent and No Use respectively. Thus, the weights assigned to each class would be 1/55.6, 1/115 and 1/2,328, and the a priori probabilities, 2.2%, 4.6% and 93.2%, respectively. The Decision Trees have also been obtained with the Rpart library of R and the values established for the cp and minsplit parameters have been 0.28 and 200, in that order.

The figure below shows the resulting Decision Tree. The colour codes and content are analogous to those shown for the use case in Leuven.

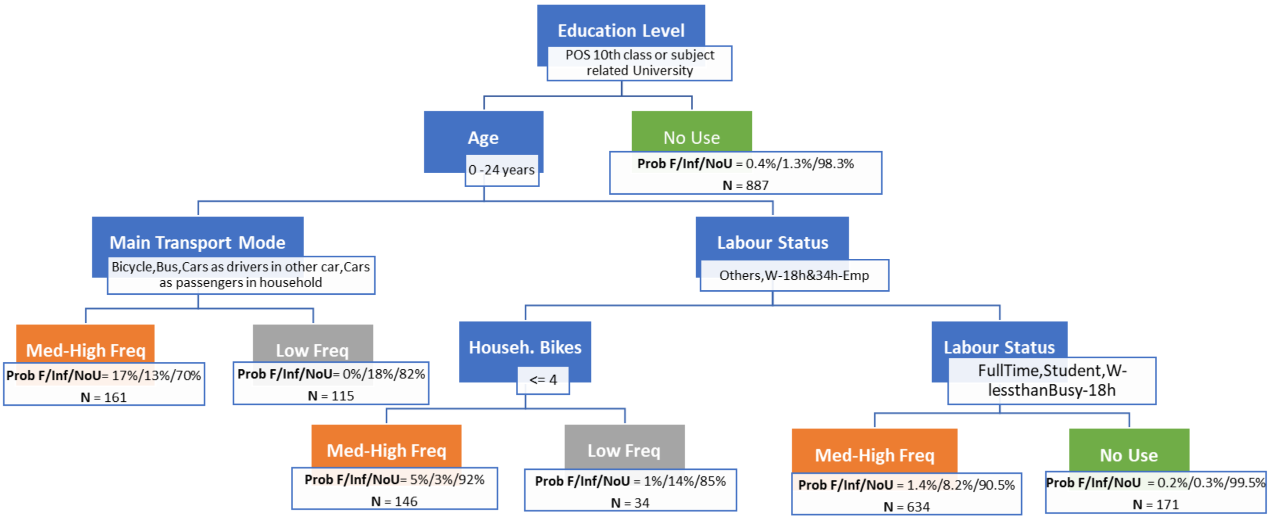


Figure 145. Decision Tree with the most relevant factors for car-sharing adoption

In the figure, we can see that the most relevant variable is the educational level since if a person does not have at least a university degree, the probability of not using car-sharing services rises to 98%. The next most relevant variable is age, and specifically the fact of being under 25. In the case of being under 25, the transport mode is the factor that best determines whether or not car-sharing is used frequently. Specifically, if this factor corresponds to Bicycle, Bus, Cars as drivers in other car or Cars as passengers in a household, then the probability of frequent and infrequent car-sharing is multiplied by 8 and 3, respectively. If the latter condition is not met, the probability of frequent use is reduced to 0, but the one of infrequent use rises to 18 per cent, i.e. it is multiplied by 4, compared to the a priori probability.

For the over-25s, labour status and the number of bicycles in the household are the factors that best predict the type of car-sharing use. In particular, if the work status is Apprentice, House-W/H, Pensioner, TExempted/onLeave or Unemployed, the probability of the individual travelling in this mode of transport is practically zero. In case the work status is FullTime, Student or W-lessthanBusy-18h, the probability of not using car-sharing is slightly reduced to 90%, while the probability of infrequent use is almost doubled (8.6%).

If the labour status is Others or W-18h&34h-Emp, then there is a higher probability of using car-sharing, and the more or less frequent use of it is mainly determined by the number of bicycles in the household. More specifically, if there are 4 or fewer bicycles in the household, then the probability of frequent use doubles (from 2.2% to 5%), while if it is greater than 4, the probability drops by half, but the probability of infrequent use increases considerably up to 14%.

3.3.2.4. Profiling of car-sharing users in Regensburg

Analogous to the use case of Leuven, in this section we intend to make a profile of the different types of users of the car-sharing service that we can find in Regensburg, taking into account the criteria analysed above such as sex, age, income, education, labour status, household size and type, as well as the number of bicycles, cars and public transport passes in the household.

The segmentation has been also carried out using the k-means method. The use of dummy and ordinal variables, as well as normalization and the value of k (number of clusters), has been done following the same criteria. In this case, the value of k has been set to five, and the number of samples discarded because of missing values were 12 out of 160. The data for each profile are displayed in Table 37,Table 38,Table 39, Table 40 and Table 41. Some of these variables are categorical and their numerical coding is presented in Table 36.

Sex	Age	Education	**Income**	
0- Male	1- 0 to 14 years	1- Apprenticeship, vocational school, Business School, School for master	1- Less than 500	6- 2.600 to less than 3.000
1- Female	2- 15 to 24 years	2- School for master craftsmen and technicians, Technical school, vocational/Specialist academy	2- 500 to less than 900	7- 3,000 to less than 3,600
	3- 25 to 44 years	3- High or University of Applied Sciences degree	3- 900 to less than 1.500	8- 3.600 to under 4.600
	4- 45 to 64 years	4- (yet) without Vocational training	4- 1.500 to less than 2.000	9- 4.600 to less than 5.600
	5- 65 years and older		5- 2.000 to less than 2.600	10- 5.600 and more

Table 36. Coding used for categorical variables in car-sharing subscriber profiling

In this case, the first user segment is shown in Table 37. As can be seen in the data, they correspond to women between 25 and 44 years old, mostly with university studies. The employment status is varied, and we can find people who work part-time (50%), study (17%), are retired (12%) or are housewives (8%). They live in two-person households without children and have medium/high incomes, so they are likely to be young couples without children. Moreover, as also observed in Leuven, these are households with few cars (one or no cars for the most part) and a high number of bicycles (2 or 3 mostly).

The second profile is mainly men, also mostly aged between 25 and 44 with university studies and working full time. Their households are similar to those of the previous group, although with a higher level of income and a higher percentage of people living alone (20%).

The third category of users corresponds to university students between 15 and 24 years of age who are women in a high percentage (70%). Most live with another person (67%) and have a low-income level. The characteristics in terms of the number of cars and bicycles in the household are similar to previous cases.

The fourth profile corresponds to men between 25 and 44 years of age, who are likely to be fathers of families with 1 or 2 children in most cases. They work full time and live in very high-income households (4,600 – 5,600 euros). The number of cars per household is still low, as in the previous cases, but the number of bicycles is higher (between 4 and 5 in most cases), in line with the higher number of members in the household.

Finally, the last profile probably corresponds to young parents between 25 and 44 years old and children under 15 as the age varies between these two categories (e.g. Median = 1 and Q3 =3). They live in households of mostly 4 people and who have a high purchasing power. The parents in this group work part-time (34-18 hours per week) while the children are mostly students. These are households with a very low number of cars since in half of the cases they do not even have a car, and a high number of bicycles (3 or more).

	Sex	Age cat	Educ Level	Inco. Cat.	Labour Status			HH Size	Household Type	HH Bikes	HH Cars	Household PT passes			
					Pensioner	Student	Weekly between 18 and 34 h employed		Multi person without children			Season ticket	Monthly	Multi-trip	Single
Mean	0.75	3.21	2.75	6.96	0.12	0.17	0.46	2.42	0.92	2.46	0.83	0.17	0.21	0.21	0.21
Std.Dev	0.44	0.83	0.53	2.03	0.34	0.38	0.51	1.06	0.28	1.56	0.92	0.38	0.41	0.41	0.41
Min	0.00	2.00	1.00	2.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q1	0.50	3.00	3.00	6.00	0.00	0.00	0.00	2.00	1.00	2.00	0.00	0.00	0.00	0.00	0.00
Median	1.00	3.00	3.00	7.00	0.00	0.00	0.00	2.00	1.00	2.00	1.00	0.00	0.00	0.00	0.00
Q3	1.00	4.00	3.00	8.00	0.00	0.00	1.00	3.00	1.00	3.00	1.50	0.00	0.00	0.00	0.00
Max	1.00	5.00	3.00	10.00	1.00	1.00	1.00	6.00	1.00	7.00	3.00	1.00	1.00	1.00	1.00

Table 37. Descriptive statistics for Group 1 (N = 24 individuals)

	Sex	Age cat	Educ Level	Inco. Cat	Labour Status	HH Size	Household Type		HH Bikes	HH Cars	Household PT passes				
					Full Time		Multi person without children	Single person under 65 Years			Daily	Season ticket	Monthly	Multi-trip	Single
Mean	0.30	3.22	2.83	7.74	1.00	2.02	0.80	0.20	2.39	0.72	0.09	0.11	0.04	0.30	0.41
Std.Dev	0.47	0.51	0.44	2.17	0.00	0.77	0.40	0.40	1.51	0.75	0.28	0.31	0.21	0.47	0.50
Min	0.00	2.00	1.00	2.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q1	0.00	3.00	3.00	6.00	1.00	2.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Median	0.00	3.00	3.00	8.00	1.00	2.00	1.00	0.00	2.00	1.00	0.00	0.00	0.00	0.00	0.00
Q3	1.00	4.00	3.00	10.00	1.00	2.00	1.00	0.00	3.00	1.00	0.00	0.00	0.00	1.00	1.00
Max	1.00	4.00	3.00	10.00	1.00	5.00	1.00	1.00	7.00	3.00	1.00	1.00	1.00	1.00	1.00

Table 38. Descriptive statistics for Group 2 (N = 46 individuals)

	Sex	Age cat	Educ Level	Inco. Cat	Labour Status		HH Size	Household Type		HH Bikes	HH Cars	Household PT passes
					Full Time	Student		Multi person without children	Single person under 65 Years			Season ticket
Mean	0.70	2.21	3.00	3.52	0.03	0.97	2.52	0.67	0.33	2.79	0.82	1.00
Std.Dev	0.47	0.42	0.00	2.22	0.17	0.17	1.64	0.48	0.48	1.88	0.95	0.00
Min	0.00	2.00	3.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00
Q1	0.00	2.00	3.00	2.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00
Median	1.00	2.00	3.00	3.00	0.00	1.00	2.00	1.00	0.00	2.00	1.00	1.00
Q3	1.00	2.00	3.00	5.00	0.00	1.00	4.00	1.00	1.00	4.00	1.00	1.00
Max	1.00	3.00	3.00	8.00	1.00	1.00	6.00	1.00	1.00	7.00	3.00	1.00

Table 39. Descriptive statistics for Group 3 (N = 33 individuals)

	Sex	Age cat	Educ Level	Inco. cat	Labour Status	HH Size	Household Type	HH Bikes	HH Cars	Household PT passes			
					Full Time		Multi person with children			Annual	Daily	Multi-trip	Single
Mean	0.12	3.35	2.88	9.24	1.00	3.71	1.00	4.65	1.06	0.12	0.12	0.12	0.24
Std.Dev	0.33	0.49	0.33	1.15	0.00	0.59	0.00	1.90	0.90	0.33	0.33	0.33	0.44
Min	0.00	3.00	2.00	6.00	1.00	3.00	1.00	2.00	0.00	0.00	0.00	0.00	0.00
Q1	0.00	3.00	3.00	9.00	1.00	3.00	1.00	4.00	0.00	0.00	0.00	0.00	0.00
Median	0.00	3.00	3.00	10.00	1.00	4.00	1.00	4.00	1.00	0.00	0.00	0.00	0.00
Q3	0.00	4.00	3.00	10.00	1.00	4.00	1.00	5.00	2.00	0.00	0.00	0.00	0.00
Max	1.00	4.00	3.00	10.00	1.00	5.00	1.00	10.00	3.00	1.00	1.00	1.00	1.00

Table 40. Descriptive statistics for Group 4 (N = 17 individuals)

	Sex	Age cat	Educ Level	Inco. Cat	Labour Status		HH Size	Household Type	HH Bikes	HH Cars	Household PT passes		
					Student	Weekly between 18 and 34 h employed		Multi person with children			Free travel auth.	Multi-trip	Single
Mean	0.64	1.96	3.57	8.11	0.43	0.39	3.93	1.00	4.68	0.68	0.14	0.32	0.32
Std.Dev	0.49	1.17	0.50	2.20	0.50	0.50	0.81	0.00	2.45	0.77	0.36	0.48	0.48
Min	0.00	1.00	3.00	3.00	0.00	0.00	2.00	1.00	0.00	0.00	0.00	0.00	0.00
Q1	0.00	1.00	3.00	7.00	0.00	0.00	3.50	1.00	3.00	0.00	0.00	0.00	0.00
Median	1.00	1.00	4.00	8.00	0.00	0.00	4.00	1.00	4.00	0.50	0.00	0.00	0.00
Q3	1.00	3.00	4.00	10.00	1.00	1.00	4.00	1.00	6.00	1.00	0.00	1.00	1.00
Max	1.00	4.00	4.00	10.00	1.00	1.00	6.00	1.00	10.00	2.00	1.00	1.00	1.00

Table 41. Descriptive statistics for Group 5 (N = 28 individuals)

3.3.3. Characterization of car-sharing trips

3.3.3.1. Applied methodologies

Using the car-sharing trip data described in Section 3.3.1, the objective of this section is to characterise these trips and extract some insights from them that allow us to understand better the use that is being made of them. Specifically, the analysis will focus on:

- The trip distances, in the sense of the distance between the origin and the destination. To this end, given that the data analysed comes from trips with the same pick-up and return station, it is assumed that the destination is at a distance equal to half the kilometres travelled. This is a heuristic rule that could lead to biases in the results, but given that there is no other information available, it is the best way to estimate the origin-destination distance.
- The duration of the trip, in the sense of the time elapsed from the picking-up of the vehicle until its return.
- The hour and the day of the week in which the vehicle is picked up and returned.
- The time in advance in which the vehicle is booked.

These variables and their interrelations are studied for both car-sharing services.

3.3.3.2. Result analysis

3.3.3.2.1. Trip distance analysis

The following figure shows the distribution of the distances of the trips made in both car-sharing services, as well as a table with the values of the main statistics.

Trip distance (km)	Mean	Std.Dev	Min	Q1	Median	Q3	Max
City	24.81	26.34	0.50	7.50	15.50	32.00	250.00
District	45.33	45.43	0.50	14.50	31.00	59.50	360.50

Table 42. Trip distance statistics in Regensburg city and district car-sharing services

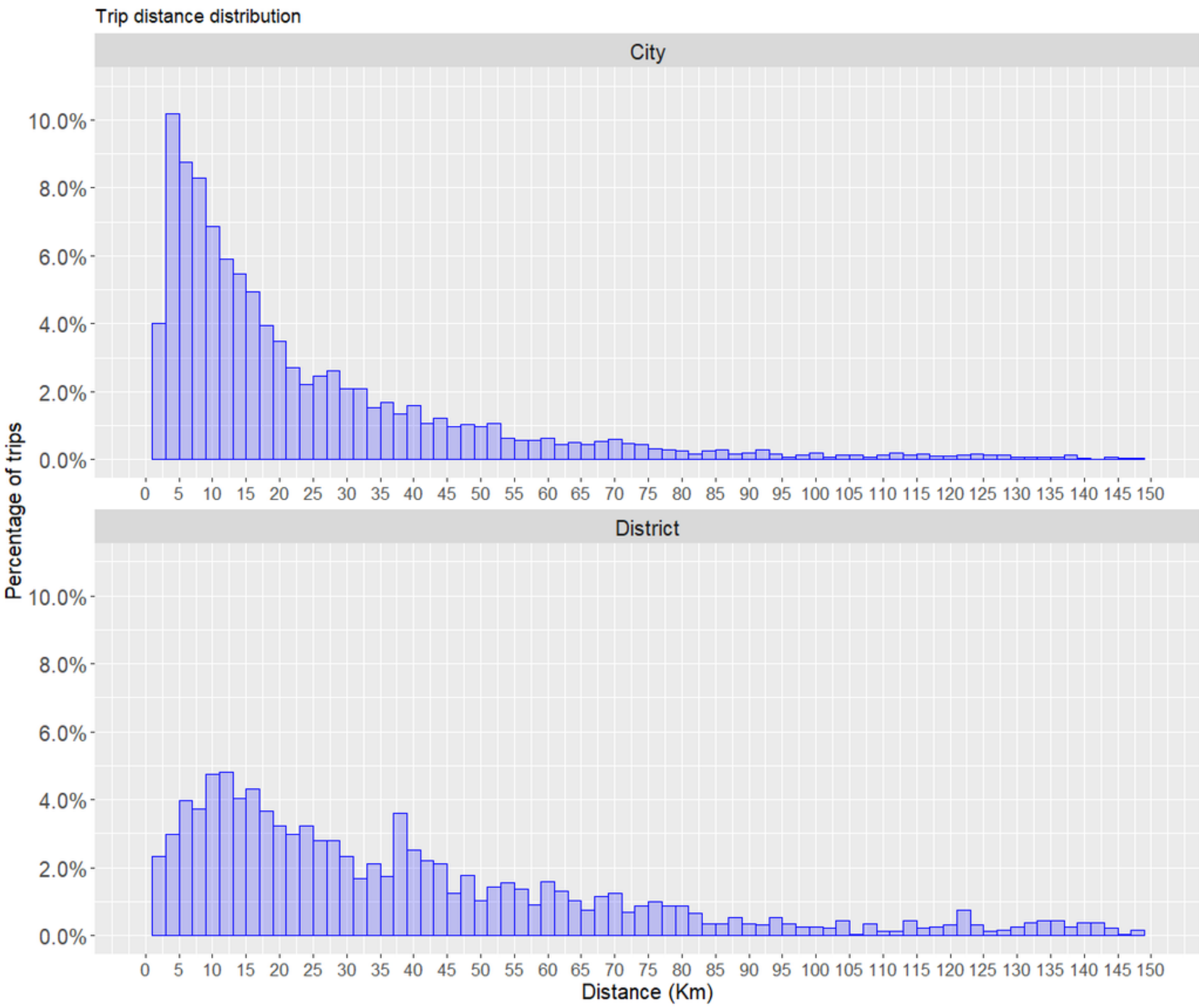


Figure 146. Trip distance distribution in Regensburg city and district car-sharing services

The mean distances of the trips (24.8 and 45.3 km, respectively) indicate that, given the area of Regensburg and the cities of its metropolitan area, they are not urban trips but medium distance ones.

However, although the distribution of the city's car-sharing service shows the existence of a great variability of distances (large tail to the right), the most frequent trip distances (mode) are in the range of 4-5 kilometres. In this case, it seems that the most frequent trips do occur within the urban area, but it represents a relatively small percentage of the total trips (10% approx.). In fact, 50% of the trips of the city car-sharing service are at most 15.5 kilometres (median).

In contrast, the distribution of the district car-sharing service is not so skewed towards such short distances, with a higher proportion of trips with medium-long distances. The mode, in this case, is 12-13km, which is the approximate distance at which many of the stations of this district service are to the Regensburg city centre, which could indicate that the most frequent trips have this destination. A second "peak" of the distribution at 37-38kms is also noteworthy here, which is probably due to travel to nearby locations at a similar distance such as Straubing.

The following tables show the main statistics of the distance distributions of the trips of both car-sharing services grouped by day of the week.

City							
Trip distance (km)	Mean	Std.Dev	Min	Q1	Median	Q3	Max
Sun	35.92	31.01	0.50	12.00	25.50	51.00	181.00
Mon	22.19	23.12	0.50	7.00	15.00	29.00	149.50
Tue	24.64	26.34	0.50	7.50	14.50	32.50	250.00
Wed	22.70	23.51	0.50	7.00	14.50	29.50	175.00
Thu	22.67	24.65	0.50	7.50	14.00	28.00	212.00
Fri	23.39	26.74	0.50	7.00	13.50	27.50	196.00
Sat	25.52	27.79	1.00	8.00	15.50	34.00	238.00

Table 43. Trip distance statistics grouped by day of the week in Regensburg city car-sharing service

District							
Trip distance (km)	Mean	Std.Dev	Min	Q1	Median	Q3	Max
Sun	53.23	40.91	2.50	24.00	43.00	70.50	295.00
Mon	43.08	48.89	1.50	13.75	26.00	52.00	311.0
Tue	37.61	40.97	1.00	12.00	23.50	42.00	276.50
Wed	42.20	47.08	1.50	12.00	27.50	51.75	360.50
Thu	39.02	40.21	1.00	13.50	26.75	49.00	331.00
Fri	48.75	51.40	0.50	15.00	31.75	59.50	313.50
Sat	52.34	44.78	0.50	18.00	43.25	69.00	293.50

Table 44. Trip distance statistics grouped by day of the week in Regensburg district car-sharing service

In the service deployed in the urban area of Regensburg, the average trip distance is similar from Monday to Saturday but shows a strong increase on Sunday. This may suggest that on that day, the service is used for leisure-related trips away from the urban area.

As for the service deployed in the Regensburg district, the average trip distance is similar from Monday to Thursday, slightly higher on Friday and noticeably higher on the weekend, with no difference between Saturday and Sunday. This may suggest that usage is different on weekends and weekdays. Weekends and perhaps Friday afternoons are reserved for trips more oriented towards leisure activities or shopping in places further away from the cities where this service is deployed.

3.3.3.2.2. Trip duration analysis

The following figure shows the distribution of the duration of the trips made in both car-sharing services, as well as a table with the values of the main statistics.

Trip duration (hours)	Mean	Std.Dev	Min	Q1	Median	Q3	Max
City	3.58	2.87	0.02	1.73	2.80	4.57	23.93
District	5.63	4.36	0.03	2.45	4.62	7.30	23.95

Table 45. Trip duration statistics in Regensburg city and district car-sharing services

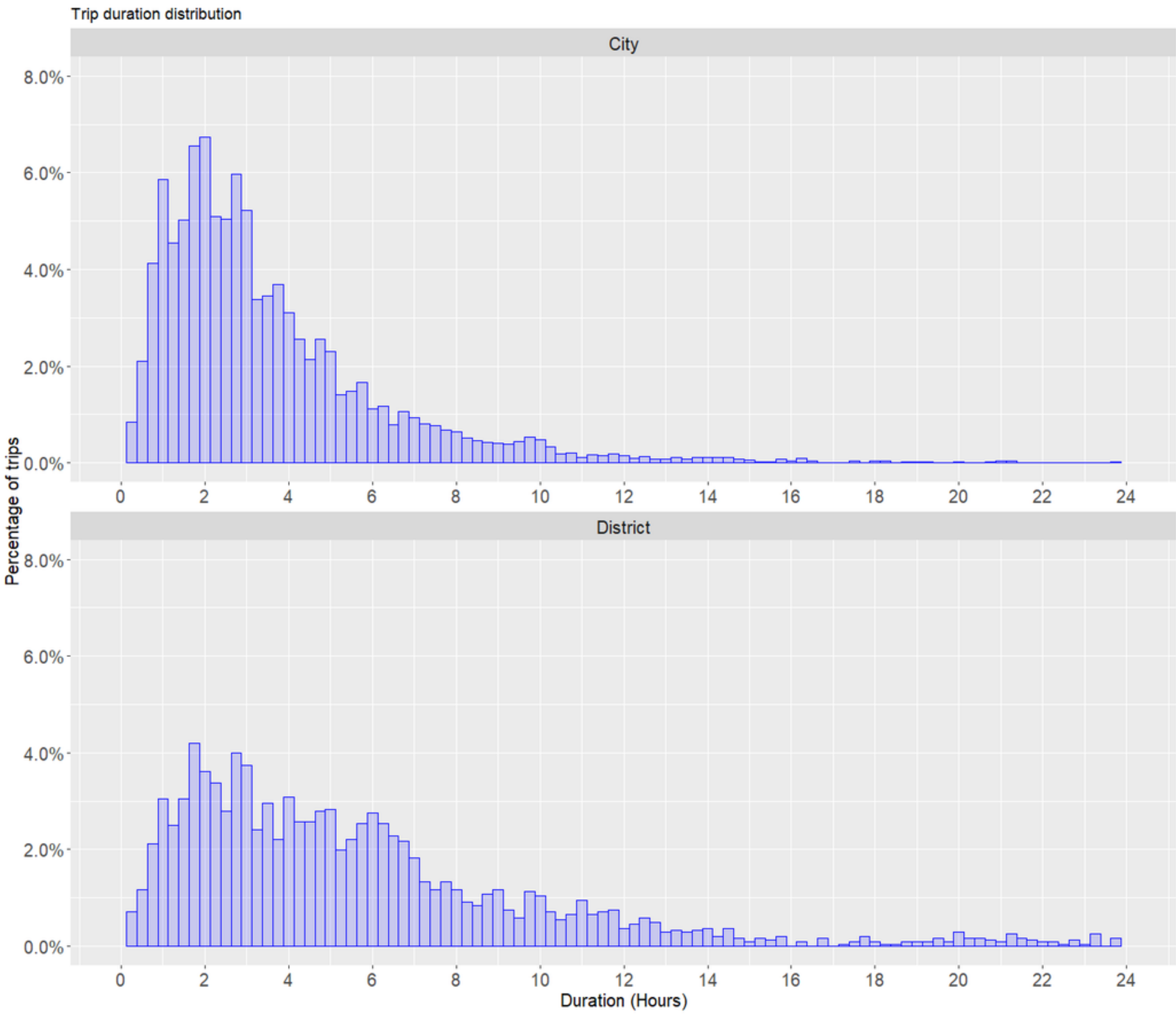


Figure 147. Trip duration distribution in Regensburg city and district car-sharing services

The average duration of trips (forward trip + activity + return trip) is probably more associated with trips related to activities such as leisure, shopping or personal issues. This possible explanation has already been considered by the scientific literature. For example, a study conducted in Switzerland (Becker, Ciari, & Axhausen, 2017) reported that most of the trips undertaken with a station-based car-sharing vehicle were shopping or leisure trips, or trips where the customer had large items to carry. In fact, when asked why they had used car-sharing for their last car-sharing ride, 40% of the members of the station-based service in the Swiss study cited goods for carrying as a main reason. In this way, the results clearly show that this is a service that is more geared towards use by the hour than by the minute. That is, it is more similar to the service provided by companies like Ubeqoo, than to that of companies like SHARE NOW. The main reason for this is the pricing scheme of the two services studied, which are charged by hourly intervals, with payment not being possible by fractions of an hour.

In the case of the service deployed in the city of Regensburg, the average duration of 3.5 hours indicates that this service is probably associated with the shortest activities among those mentioned above, such as shopping, errands or personal issues.

In contrast, the average duration for the car-sharing service in the district is significantly longer, specifically, 5.63 hours. Although the average distance was also greater, the increase in time is not justified by the increase in distance (about 40 kilometres more on average counting the round trip) so they are probably associated with longer duration activities. In any case, it is also observed that the third quantile is close to 8 hours, so there may also be a certain percentage of trips associated with work activities.

Analysing the shape of the distributions, it can be seen that the service in the city is more biased towards the left than the service in the district, as was already the case with the distance of the trips. However, this bias is less pronounced in the case of trip distance. The most frequent duration for both services is 2 hours (mode), with 3 hours also being common. It is interesting to see how some peaks appear at exact or nearly exact hours in both services. This is probably due to the pricing systems applied, since reservations are paid by blocks of hours regardless of whether they are used in full or not.

The following tables show the main statistics of the trip duration distribution of both car-sharing services grouped by day of the week.

City							
Trip duration (hours)	Mean	Std.Dev	Min	Q1	Median	Q3	Max
Sun	4.31	3.13	0.08	2.00	3.57	5.72	21.32
Mon	3.40	2.63	0.13	1.72	2.72	4.42	23.88
Tue	3.46	2.87	0.07	1.73	2.70	4.13	20.97
Wed	3.37	2.64	0.15	1.62	2.72	4.25	21.28
Thu	3.41	2.82	0.02	1.60	2.67	4.20	23.83
Fri	3.47	2.93	0.22	1.72	2.70	4.22	23.93
Sat	3.90	3.03	0.22	1.82	3.01	4.98	23.28

Table 46. Trip duration statistics grouped by day of the week in Regensburg city car-sharing service

District							
Trip duration (hours)	Mean	Std.Dev	Min	Q1	Median	Q3	Max
Sun	5.67	3.80	0.40	2.92	5.20	7.25	23.25
Mon	5.42	4.60	0.08	2.00	4.05	7.24	23.85
Tue	5.56	4.62	0.10	2.37	4.12	6.78	23.95
Wed	5.17	4.25	0.25	2.17	3.98	6.57	23.25
Thu	5.00	3.84	0.17	2.20	4.05	6.41	21.58
Fri	5.92	4.63	0.18	2.65	4.83	7.80	23.78
Sat	6.61	4.58	0.03	3.30	5.65	8.67	23.85

Table 47. Trip duration statistics grouped by day of the week in Regensburg district car-sharing service

In the city car-sharing service it is observed that longer trips are made during the weekend, especially on Sunday. Meanwhile, the working days are quite similar among them, with trip durations slightly lower than 3.5 hours. This data is in line with what was previously commented, in the sense that the city service is probably used for shopping or for personal issues and, only on Sundays, leisure activities are more frequent.

On the other side, the service for the district shows a slightly longer average trip duration on Saturdays and Fridays, but not on Sundays. In addition, the third quantile of Saturdays is above 8 hours, which may indicate a higher percentage of work-related trips. The reason may be less frequent or lower availability of public transportation connections on weekends compared to weekdays.

3.3.3.2.3. Trips per time of the day analysis

The following figure shows the number of trips according to their start time (hour) and day of the week, for both car-sharing services.

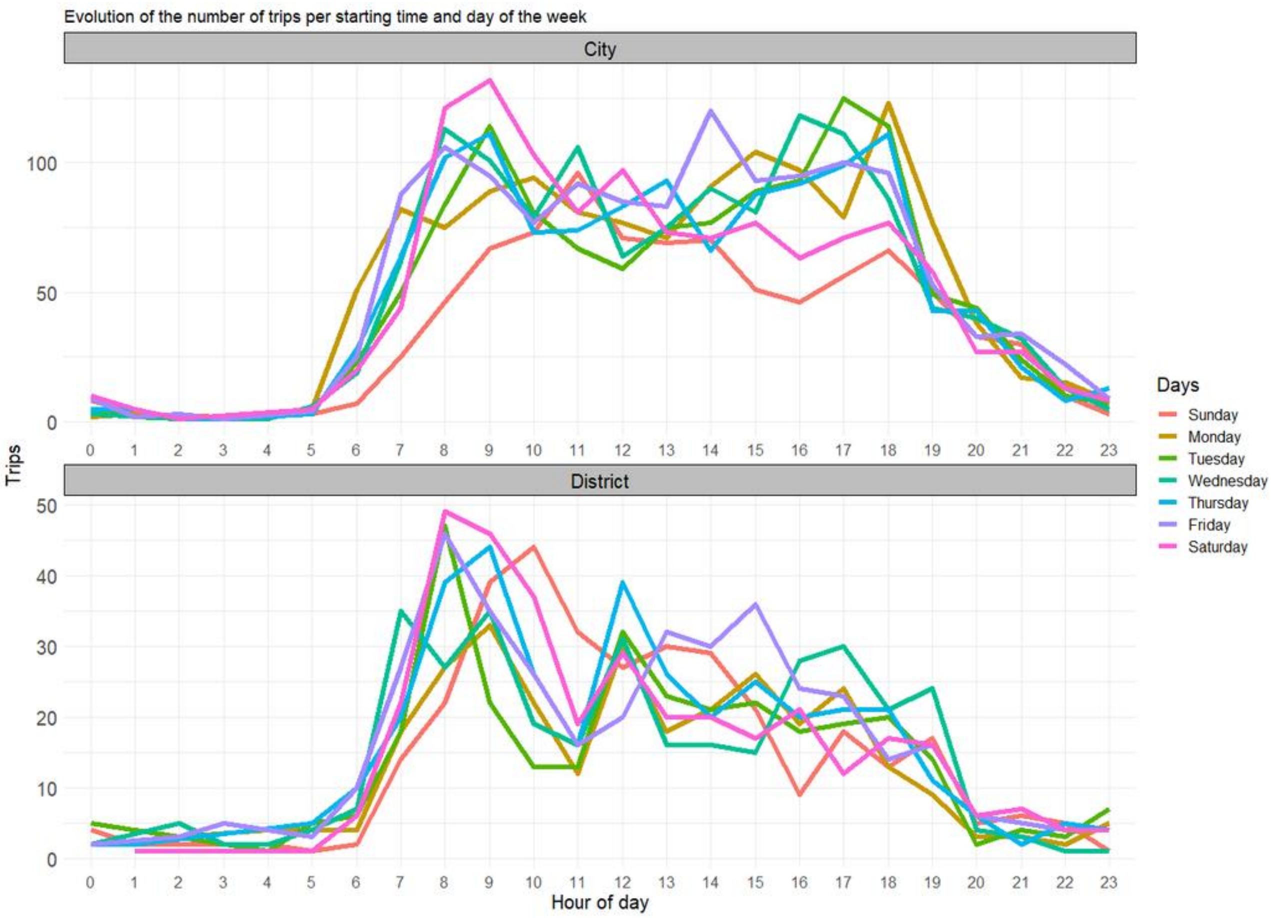


Figure 148. Number of trips according by the start time and day of week in Regensburg city and district car-sharing services

As for the service in the city, it is observed that during the weekdays there are two peaks for the start of the trips: the first one in the afternoon and the second one in the morning. The one in the afternoon is between 4pm and 6 pm from Monday to Thursday, while on Fridays it is earlier, concretely, at 2 pm. In the morning, the peak happens between 8am and 9am from Tuesday to Friday, being delayed one hour on Mondays. During the weekends, there are also two peaks but the main one is at 9am, which is delayed until 11am on Sundays. The afternoon peak, much less pronounced, occurs around 6pm.

As for the service in the district, on most days there is only one peak in the morning, between 8am and 9am in the morning (10am in the case of Sundays). In addition, from Monday to Saturday there is a secondary peak around 12, which is delayed until 3pm on Fridays.

Similarly, the distribution of trips according to their end or return time (time at which the car is returned to the station in which it was picked-up) is shown below.

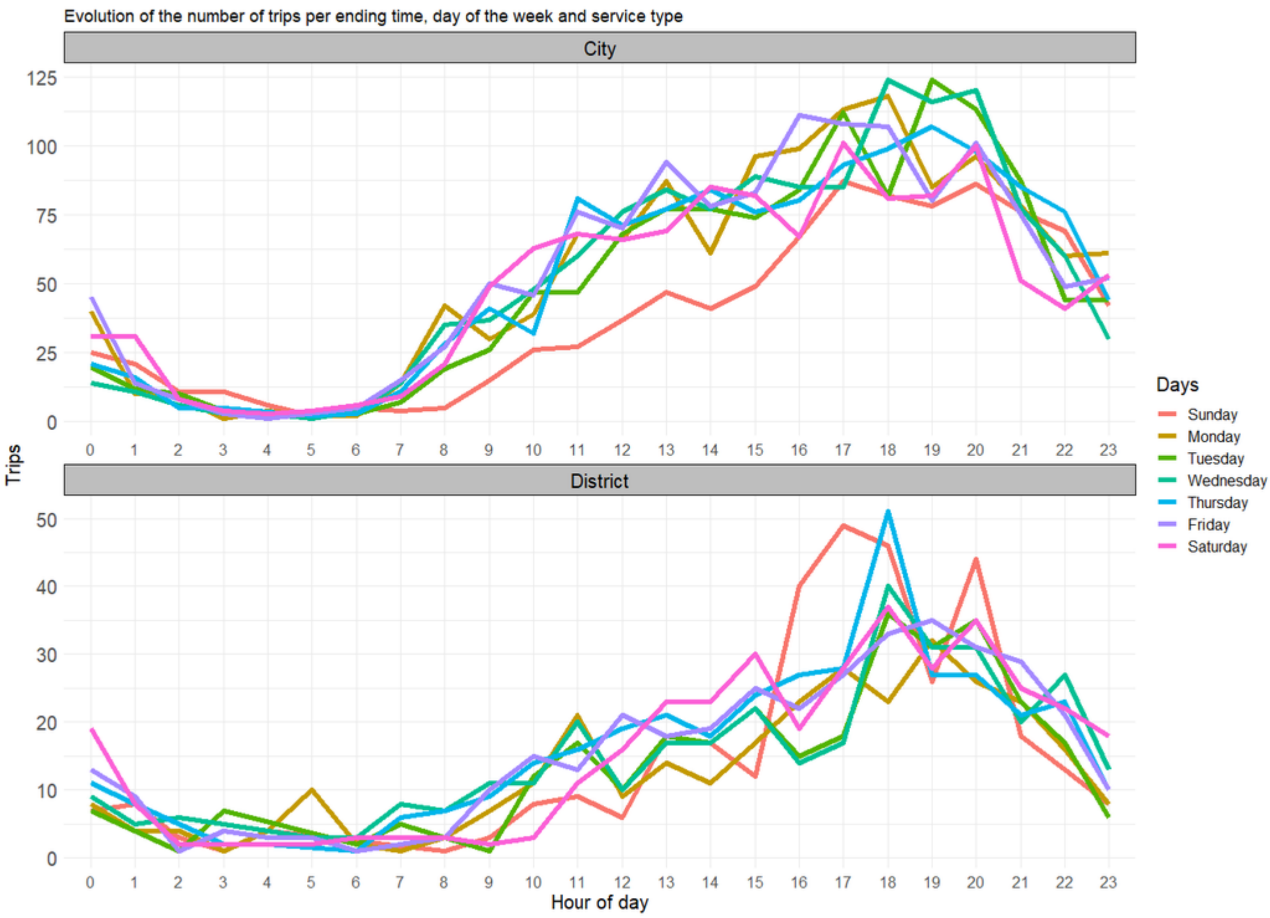


Figure 149. Number of trips according by the end time and day of week in Regensburg city and district car-sharing services

In the city service, it is observed that number of trips endings (e.g., vehicles returned to their corresponding stations) grows steadily as the day progresses, reaching its peak between 6 and 8 pm, moving forward to 4 pm on Fridays. From that moment on, returns clearly decrease. For the district service it happens in a similar way, although on Sundays the peak tends to be earlier, at around 5pm.

Below are the distance distributions of the trips grouped according to their start time and the type of day (weekday/ weekend), for both car-sharing services.

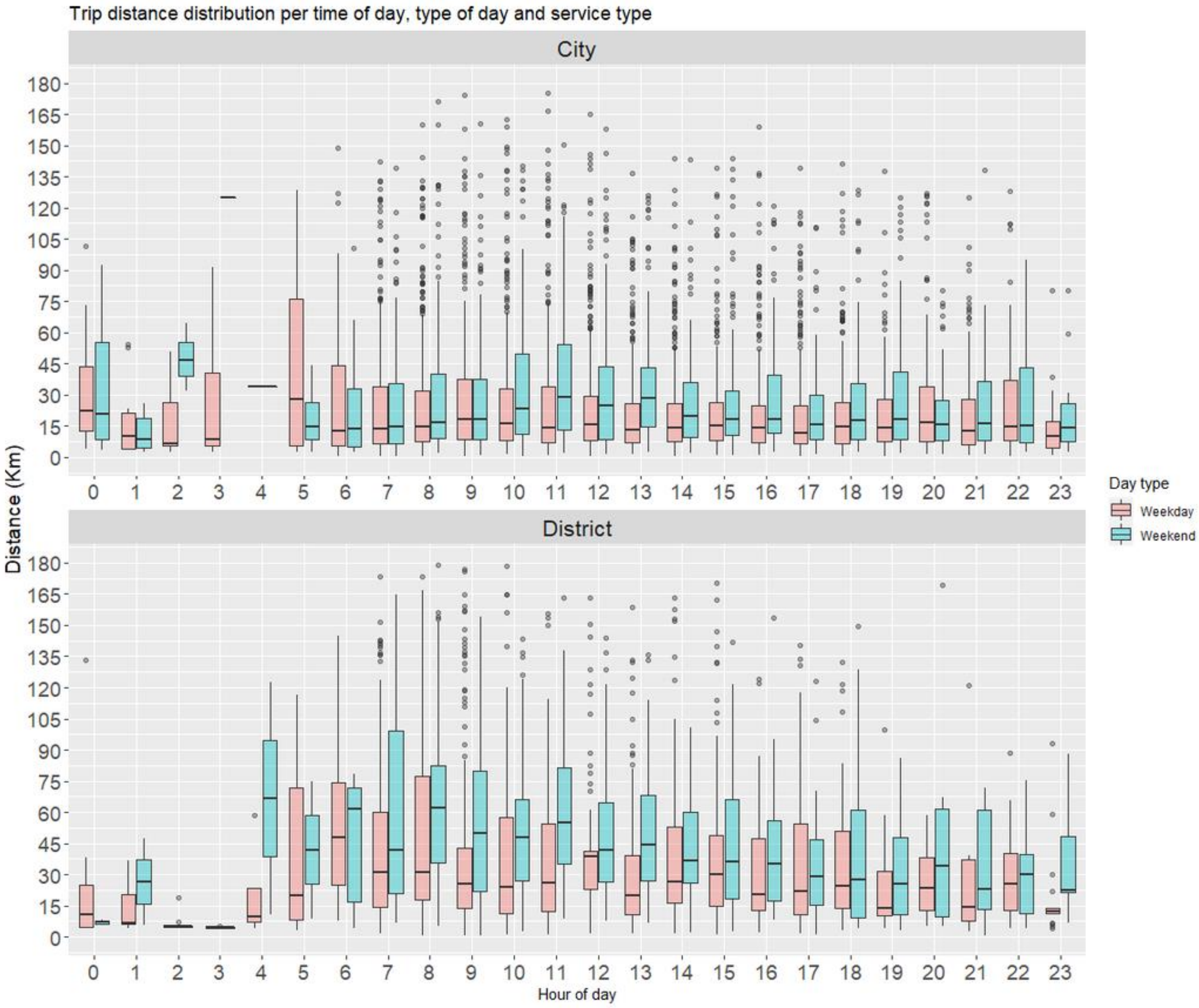


Figure 150. Distribution of trip distance according by the start time and type of day in Regensburg city and district car-sharing services

As for the city service, during weekdays, the longest trips are concentrated in the early morning hours (5 and 6). If we look at the average (median) we see that (ignoring data from night trips because they are noisier) it rises slightly from 6 to 9, then drops slightly to rise again around 6pm and drops again at 11 at night. On the weekends the behaviour is different, since that distance peak takes place at 11 in the morning (arriving at 30km), then to descend slightly and to stabilize around the 15km from the 15 hours. The difference between weekdays and weekends is again clearly seen in this graph, with longer trips and later hours during the weekend.

On the district service side, and again avoiding trips in the night because they present noisy data, we see that, during weekdays, the longest trips occur around 8 am, although a peak is also observed at 12pm. On weekends the trend is clearer: the longest trips on average occur around 8 am (60 km) and then gradually decrease to 30-40 km.

The next and last figure in this section is similar to the previous one but relates to the temporary duration of the trips.

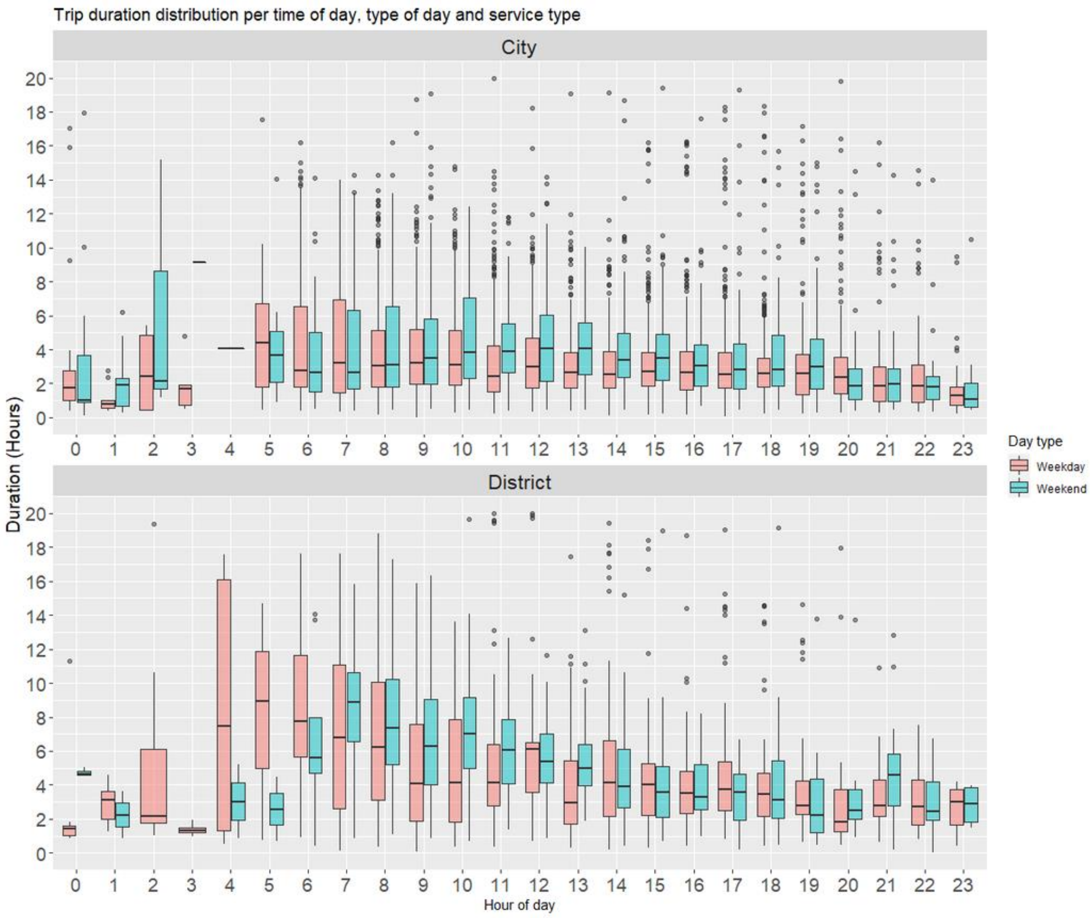


Figure 151. Distribution of trip duration according by the start time and type of day in Regensburg city and district car-sharing services

With respect to the city service, it is observed that on weekdays, the longest trips are those that begin first thing in the morning, between 5am and 7am, with an important proportion of trips with durations longer than six/seven hours. Then, as the day progresses, the durations decrease slowly but progressively. In the case of weekends, it can be seen that the longest trips do not occur as early as on weekdays but around midday (from 11am to 1pm).

And, as far as the district service is concerned, we clearly observe the longer general duration of the trips compared to the city service. For weekdays, there is a significant peak in the early hours of the day, between 5am and 8am. On the other side, there is also a peak during the weekends, but it occurs later, between 7am and 10am. In addition, weekends present longer trip duration for the rest of the morning hours and until early afternoon.

Analysing the results offered by the four graphs, the following conclusions can be drawn:

- Car-sharing trips that start in the morning are usually longer than those in the afternoon, as can be seen from the distributions and the fact that the vehicles are returned mainly at the end of the afternoon. This can be interpreted as meaning that depending on the time of the start of the trip, the activity to be carried out is different.
- The city car-sharing service is used during the week to perform short tasks in the afternoon, possibly shopping or personal issues after leaving work or class.
- Oppositely, the district service seems to be more associated with work activities, since an important percentage of them start early in the morning and have longer distances and durations.

- In both services, there is a clear difference between weekdays and weekends, with longer trips and later hours during the weekend, possibly associated with leisure activities.
- Early on Saturdays, there is a level of car-sharing use which, due to the distance and duration of the trips, does not seem to be related to leisure activities. This may point to the fact that on weekends, specifically Saturdays, car sharing may be used for trips to work to locations with restricted accessibility by public transport then.

3.3.3.2.4. Trip booking analysis

Finally, the degree of spontaneity with which the car-sharing services are used is analysed from the point of view of the time that the car is booked in advance (time-lapsed between the booking of the trip and the starting time of the trip). The following graph shows the distribution of the time in advance that the trip is booked (in hours), for both car-sharing services.

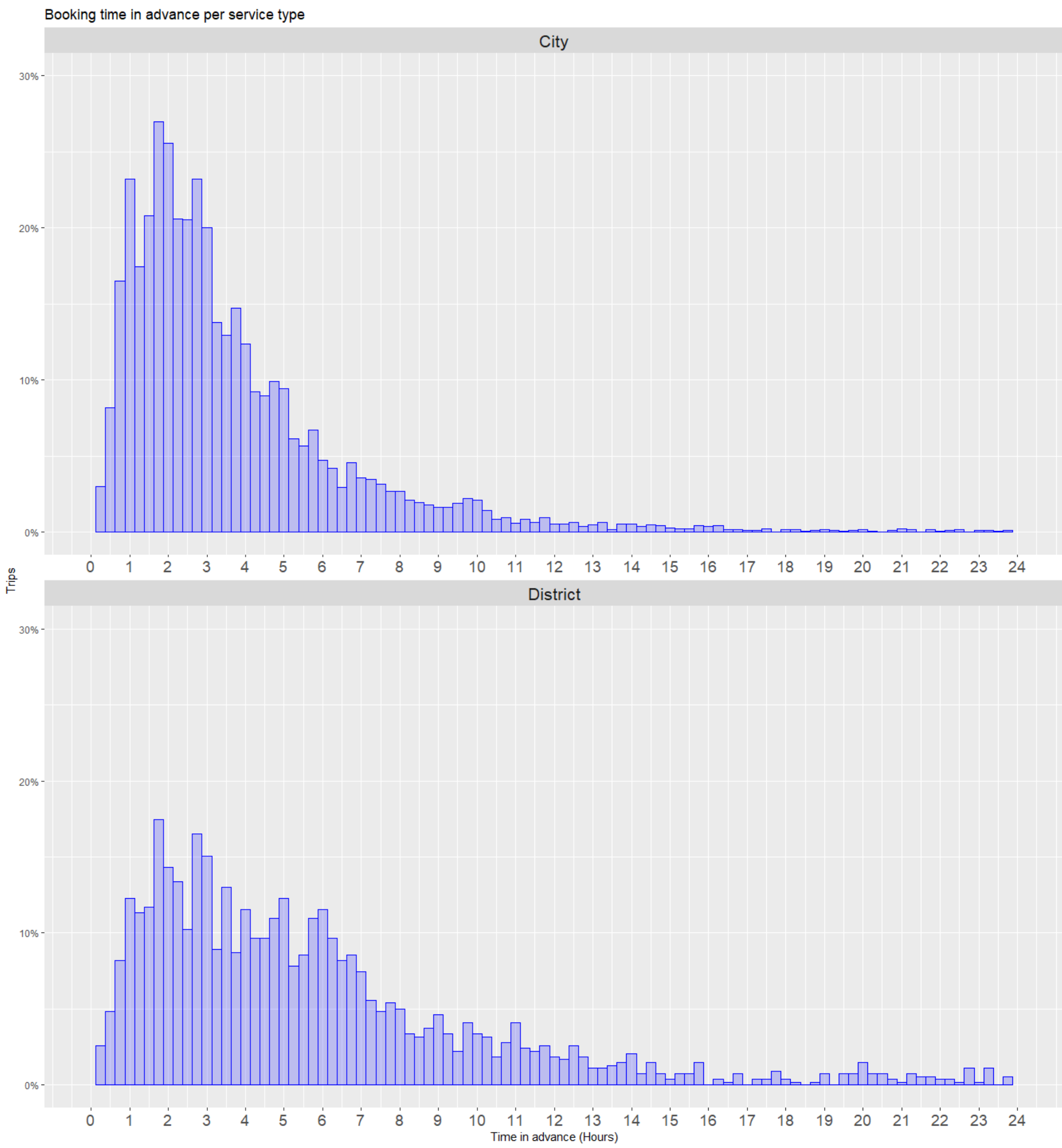


Figure 152. Distribution of the time of booking in advance (in hours) for Regensburg city and district car-sharing services

For the city service, it can be seen that the most common reservation time (mode) is 2 hours, with 50% of the trips being booked with less than 5 hours in advance, and 75% less than 7 hours. On the contrary, for the district service, the distribution is less biased towards the left. Although the most common time in advance booking is also two hours, the distribution is flatter, indicating that this service is used less spontaneously than the city service. In any case, it is not a planning made with much anticipation since 75% of the trips are made with 9 hours or less, being residual the reservations beyond 13-14 hours.

The following graph shows the distribution of trip bookings but grouped by time of day and type of day (weekday/weekend).

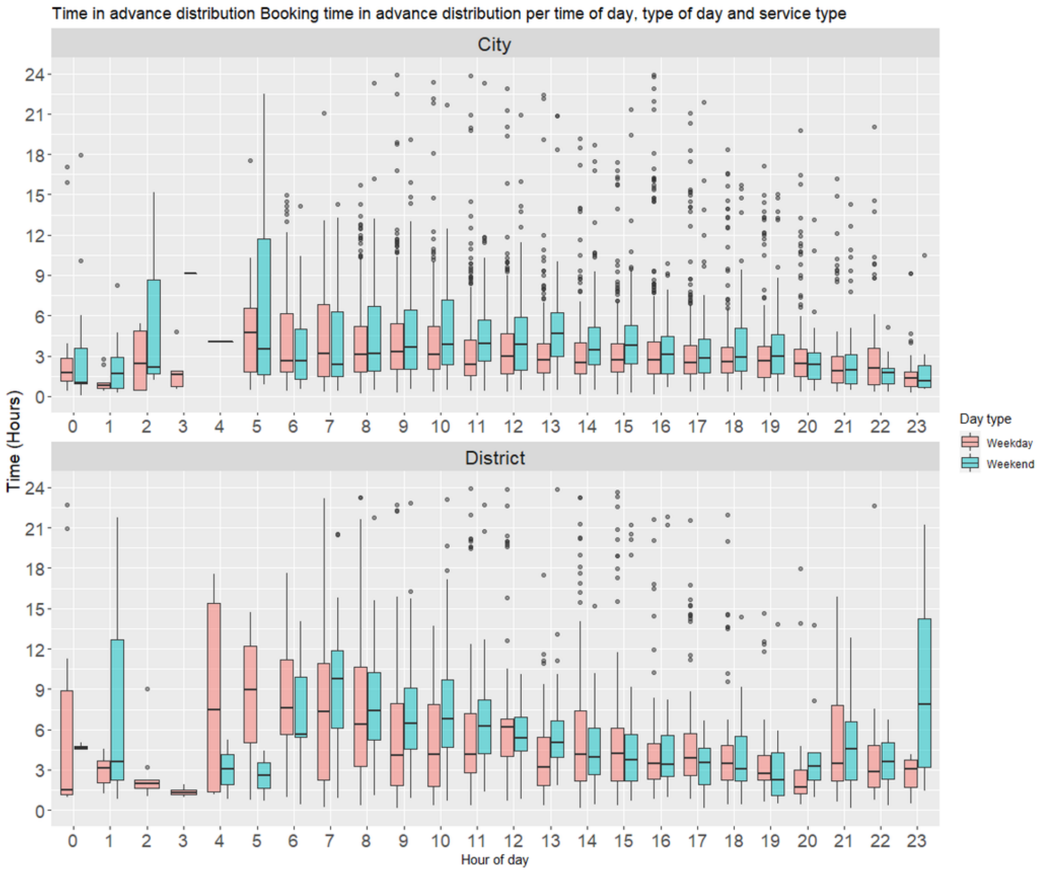


Figure 153. Distribution of the time of booking in advance according to the time of the day and type of day for Regensburg city and district car-sharing services

With respect to the city service, it can be seen that, during the week, trips that begin early in the day (from 6 to 10 AM) are planned further in advance than those at the end of the day. This seems to indicate that the use of car-sharing in the afternoon is done spontaneously, possibly related to activities such as shopping or personal issues, while morning trips may be more related to more planned activities. At the same time, weekends present a significant increase in reservation time, possibly because of the higher demand on Saturdays for leisure activities.

As for district service, the trends are similar: more advanced bookings for trips in the early hours of the day and especially at weekend. In addition, the time in advance booking for this service is significantly higher than that of the city service, possibly because the lower supply (only one car-sharing station per town/village) and because there are fewer possibilities of covering the need with public transport. Finally, it is worth noting the high values of the time in advance for booking reached by some weekday trips between 7 and 8 a.m., which may suggest a possible relationship with professional or work trips.

These results related to the time in advance for booking with respect to the beginning of the trip do not fit with what is reported by some works in the literature. Although it is not a frequently analysed variable, a study conducted in Switzerland that compared station-based and station-free car-sharing services (Becker, Ciari, & Axhausen, 2017) found that 62% of the station-based car-sharing members had planned their last trip at least one day ahead. In contrast, 72% of free-floating members planned their car-sharing trip less than one hour in advance. This may be related to a lower demand for these two services compared with those analysed in literature which make it possible to book the trip with less time in advance. But this is an aspect that should be further investigated.

3.3.3.3. Comparison of car-sharing trips versus general mobility

The aim of this sub-section is to compare car-sharing trips with general mobility to try to get a better insight into the types of trips for which this service is being used and with which other transport modes it would be competing. To do this we have used trip data from the 2017 Regensburg household survey and compared it with the car-sharing service deployed in the city of Regensburg. The main reason for considering only the city service is that the household survey only collects trips of people living in the urban area of this city, and therefore the trips are not comparable with those made by the service deployed in the district of Regensburg. Furthermore, in order to make this comparison more accurate, we have used only those car-sharing trips made on Tuesday, Wednesday or Thursday, as these are the days on which the survey was carried out. In total, there were 4698 trips. On the other hand, from the household survey we discarded those trips with no valid route information, length, and purpose, so a total of 6900 trips were analysed.

To begin this analysis, we will first compare the trip distance distribution of the general mobility in Regensburg (e.g., considering all modes of transport and all travel purposes) and that of the car-sharing trips we saw above. As shown in Figure 154, car-sharing service is used for longer and less frequent trips than the ones for which other modes of transportation are used.

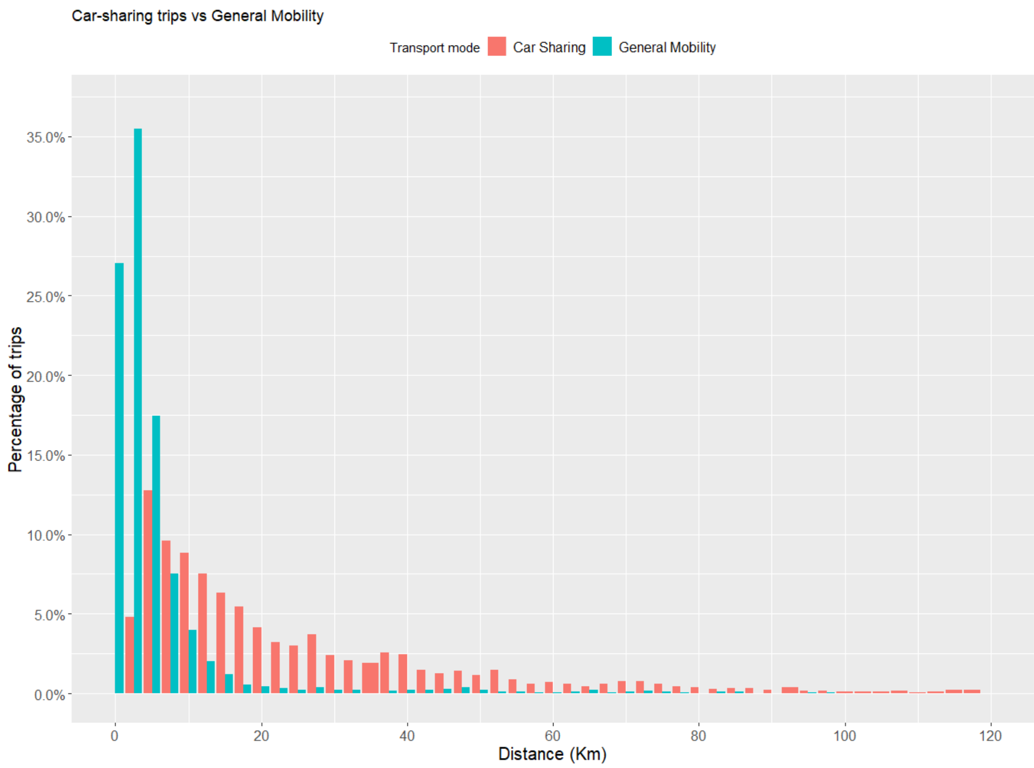


Figure 154. Trip distance distributions from Regensburg's city car-sharing service and the general mobility information in the 2017 household survey

Disaggregating the general mobility information obtained from the 2017 household survey into the three most common modes of transport in Regensburg (bus, car as driver in household car and car as Passenger in household car), it is observed that the distance of trips made with the car-sharing service has a distribution more similar to that of the two modes using the car than to that of the bus, as shown in Figure 155. In addition, a larger right-hand tail is observed, which again confirms a greater use of car-sharing for longer trips than is usually the case with the other usual modes of transport.

This greater similarity with the use of the private car suggests greater competition with this mode of transport than with public transport (in particular, the bus). However, with these data, it is likely that car-sharing is also stealing some trips to the bus over distances of less than 10kms, although, given the difference in volume for short distance trips, such trip capture is likely to be negligible.

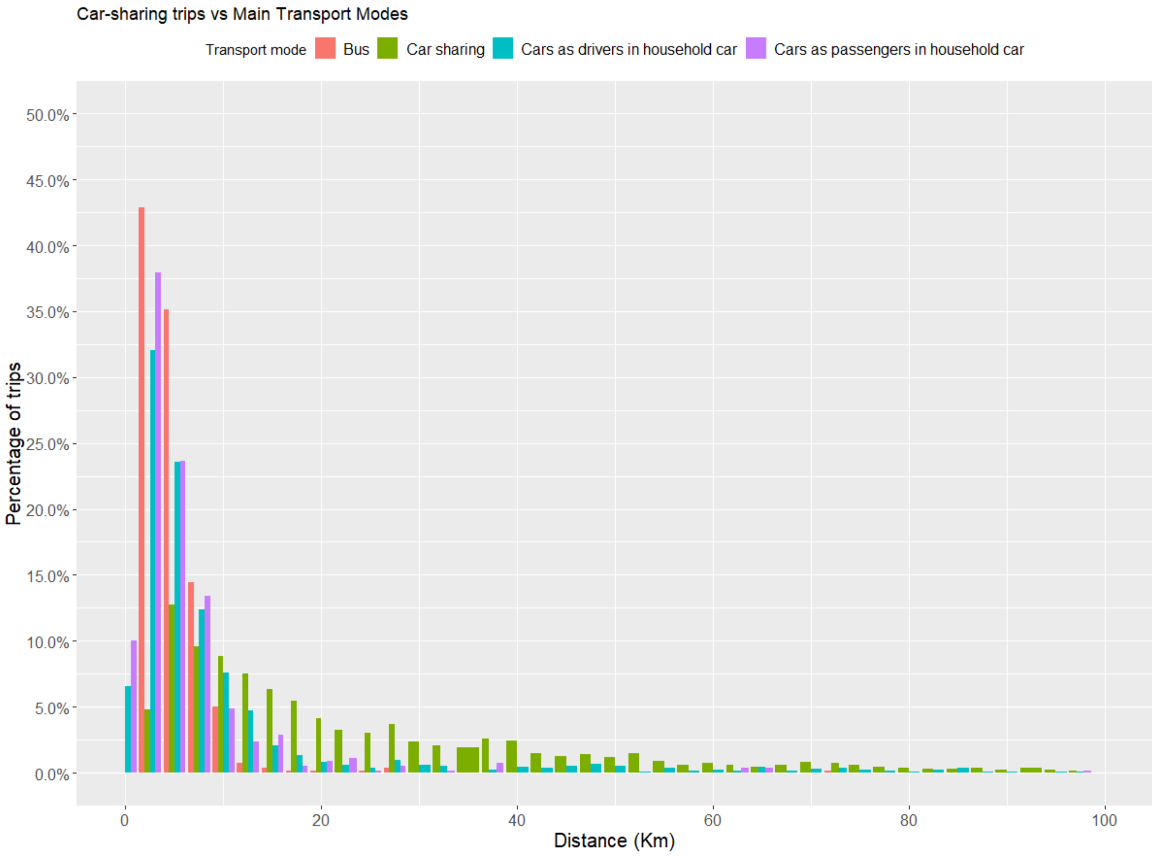


Figure 155. Trip distance distributions from Regensburg's city car-sharing service and the main transport modes identified in the 2017 household survey

Figure 156 shows a comparison of the evolution of the volume of trips throughout the day for the same four modes of transport studied. Once again, the car-sharing service is used for less frequent trips, since its profile is very different from that of the other three modes of transport analysed. However, these differences are greater when compared with the bus than when compared with the use of the private car, again suggesting less competition with the first mode of transport.

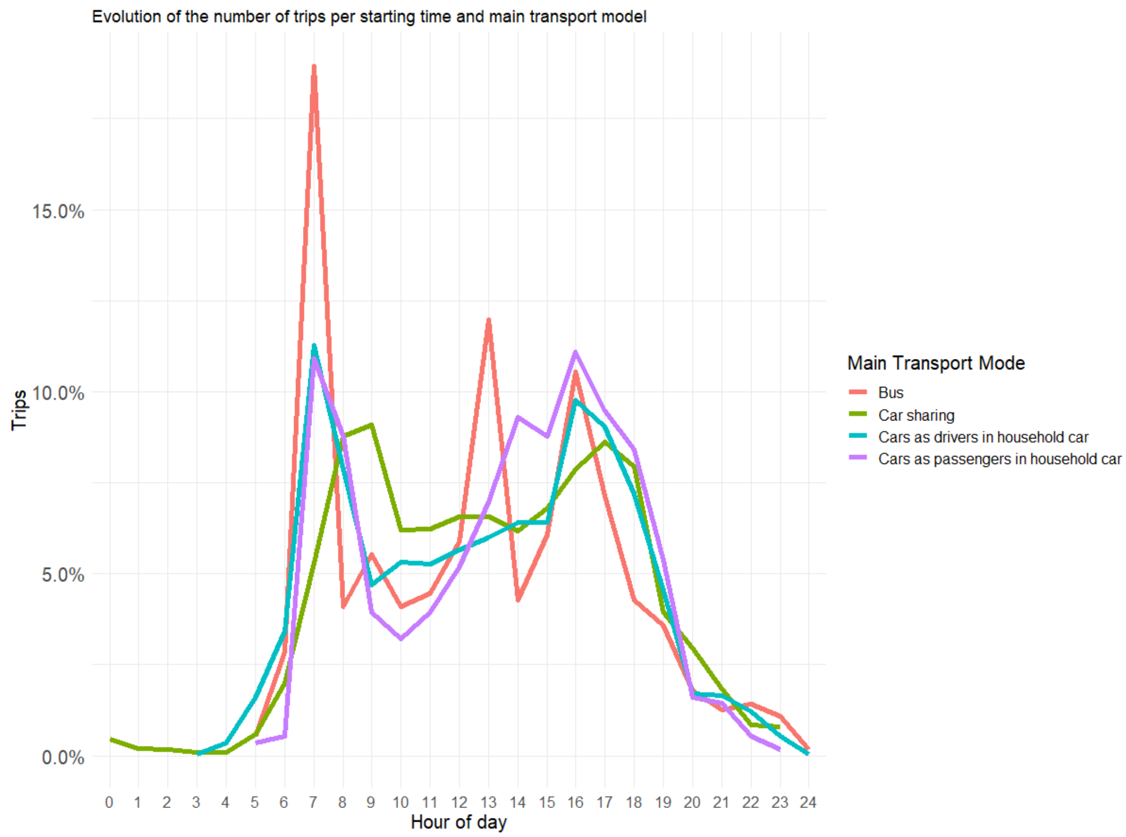


Figure 156. Distribution of number of trips per starting time for the Regensburg's city car-sharing service and the main transport modes identified in the 2017 household survey

Finally, Figure 157 shows the evolution of the volume of trips throughout the day for different travel purposes, based on data from the 2017 household survey, and is compared with the daily evolution of trips made with the car-sharing service. This can help us to understand what the main activities related to the use of car-sharing are, at least on the central days of the week.

During the first half of the day (e.g., before 12pm), we see that the profile of car-sharing trips presents a greater similarity with trips related to shopping and personal matters. In the second half of the day, the similarity is again high with respect to shopping activities, but also with respect to leisure activities. The similarity to travel related to personal matters disappears, especially after the end of the afternoon.

This confirms some of the hypotheses launched previously in which we linked the use of car-sharing to activities such as shopping, personal matters or leisure. These results are in line with the results shown in other works, such as (Becker, Ciari, & Axhausen, 2017).

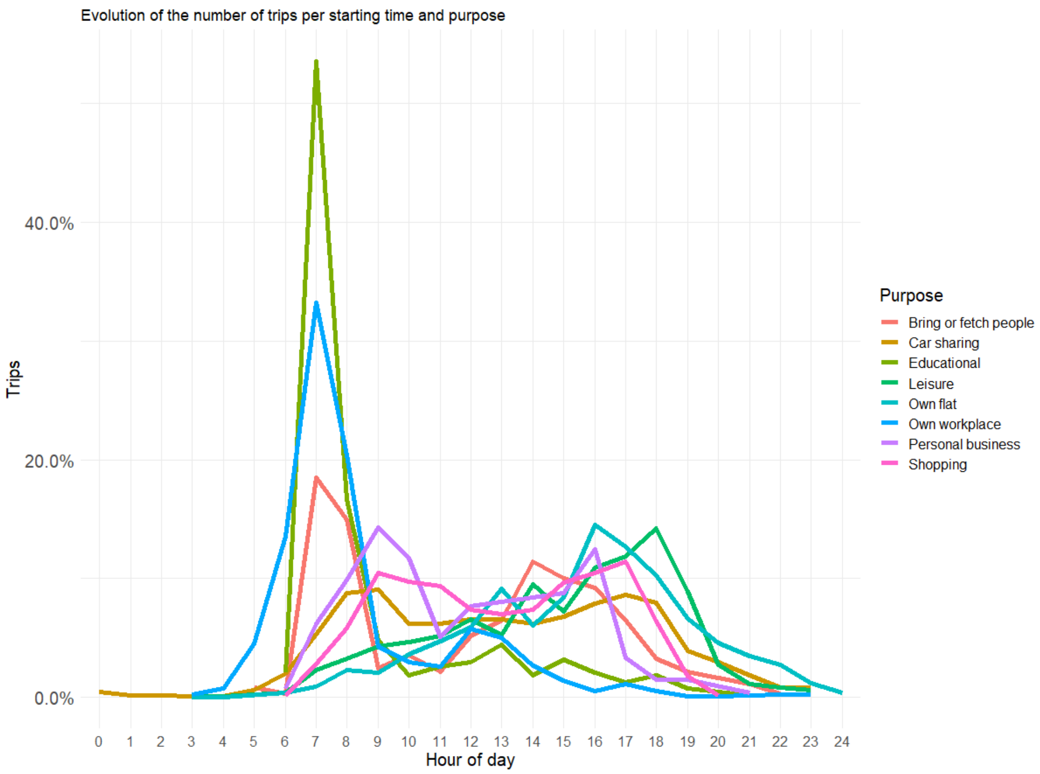


Figure 157. Distribution of the number of trips per starting time for the Regensburg's city car-sharing service and the 2017 household survey (grouped by trip purpose)

3.3.3.4. Profiling of car-sharing trips in Regensburg

The objective of this section is similar to the other profiling we did in sections X and X, although with the main difference that on this occasion, we sought to segment trips and not people. With this segmentation we try to determine which are the "typical" or "standard" trips that are usually made with this type of service, differentiating between those made in the service deployed in the city and in the district of Regensburg. The criteria we have used for the categorisation of trips are whether the trip is made on a weekday or a weekend, the start time, the time in advance for booking the trip, the duration of the trip and the distance of the trip (i.e., total distance divided by two to keep consistency with previous analysis).

The segmentation has also been carried out using the k-means method. The use of dummy and ordinal variables have not been necessary in this case since all variables were numeric or binary. We remove outliers by removing trips with values below and above percentiles 5 and 95, in trip distance, duration or time in advance for booking. Apart from this, we also applied normalization. The value of k was set using the elbow law, and the specific values were 8 for City service and 7 for District service. The data for the profiles of the City services are displayed in Tables Table 48, Table 49, Table 50, Table 51, Table 52, Table 53, Table 54 and Table 55, and the ones for the District service in Tables Table 56, Table 57, Table 58, Table 59, Table 60, Table 61 and Table 62.

Starting with the City car-sharing service, we will first analyse the trips made during weekdays, which correspond to the first six groups. In the first category, we have trips planned well in advance, that are booked 60 days before the starting of the trip on average. In addition, they show a high variance in the starting time, although most of them (75%) are done during the afternoon. They have an average duration and distance of about 3 hours and 20 kilometres, although the latter also with significant variance. This group could be related to personal issues or business trips known well in advance, but the high variance in starting time and distance could also indicate a high variety in trip purposes.

The second category corresponds to trips that take place in the morning (between 8:30 and 12:30 mainly), with shorter duration and distance than the previous group. Specifically, about 2 hours and 12 kilometres, respectively. Since they seem to be more spontaneous trips (they are booked about 2 hours in advance (median)) their purpose can be related to activities such as shopping or personal issues.

The third group contains trips similar to the previous group but made in the afternoon. Therefore, it also seems that it may be related to activities such as shopping or leisure, mainly because of the starting time.

The fourth cluster groups trips that take place mainly during normal office hours (9-16). In addition, they have a longer duration and distance than the previous groups (more than 4 hours and 50 kilometres) and are also booked further in advance (50% of them with more than 4 hours). Therefore, this cluster could correspond to business trips or personal commitments to locations or sites around Regensburg.

The fifth category of trips seems similar to the previous group, but their trips start a little earlier in the morning (between 9 and 13), are also booked a little earlier (median 6 hours), last longer (6 hours on average) but cover a shorter distance (about 20 kilometres on average). Therefore, they may also be related to business trips but also to personal issues.

The last group belonging to the weekdays contains trips that start during the afternoon (around 17:00). Since their duration, distance and advance booking are not very large (a little more than 3 hours, almost 30 kilometres and about 3 hours in advance, respectively), it seems that this group could be related to spontaneous activities carried out after the working hours, such as shopping or leisure activities.

As for the clusters related to the weekend, the first one corresponds to longer trips: 41 kilometres and 5 hours on average. Given that the start time of the trips is around noon and that they are booked about 5 hours in advance, this group could belong to leisure activities.

Finally, the second group for weekends corresponds to more spontaneous and shorter trips, which are booked about 2 hours in advance, have an average duration of 2.5 hours and an average distance of 14 kms. These data show that they can be mainly shopping trips.

	Start time	Booked in advance (h)	Duration (h)	Trip Distance (km)
Mean	14:15	1,473.88	3,2	22,5
Std.Dev	4:48	331.15	1,6	13,7
Min	1:54	850.00	0,8	5,0
Q1	12:30	1,229.00	2,0	10,5
Median	14:09	1,542.00	3,0	18,5
Q3	18:34	1,739.00	4,0	31,0
Max	23:27	1,955.00	7,8	62,0

Table 48. Statistics of the group 1 (N=123) of trips of the Regensburg city car-sharing service

	Start time	Booked in advance (h)	Duration (h)	Trip Distance (km)
Mean	10:49	39.32	2,3	12,3
Std.Dev	2:08	140.73	0,9	7,0
Min	0:01	1.00	0,8	4,5
Q1	8:27	1.00	1,7	7,0
Median	10:47	2.00	2,2	10,5
Q3	12:33	3.00	3,0	15,5
Max	14:32	879.00	4,8	44,0

Table 49. Statistics of the group 2 (N=1,245) of trips of the Regensburg city car-sharing service

	Start time	Booked in advance (h)	Duration (h)	Trip Distance (km)
Mean	17:02	35.78	2,3	10,8
Std.Dev	2:10	130.64	0,8	4,6
Min	13:15	1.00	0,9	4,5
Q1	15:18	1.00	1,7	7,0
Median	17:59	2.00	2,2	10,0
Q3	18:22	2.00	2,8	14,0
Max	24:47	889.00	5,6	24,5

Table 50. Statistics of the group 3 (N=1,477) of trips of the Regensburg city car-sharing service

	Start time	Booked in advance (h)	Duration (h)	Trip Distance (km)
Mean	12:14	46.65	4,7	53,1
Std.Dev	4:08	152.15	1,6	9,8
Min	0:04	1.00	0,8	34,0
Q1	9:08	3.00	3,6	45,5
Median	12:36	4.00	4,5	52,0
Q3	15:58	6.00	5,7	61,0
Max	23:54	1115.00	9,2	72,0

Table 51. Statistics of the group 4 (N=491) of trips of the Regensburg city car-sharing service

	Start time	Booked in advance (h)	Duration (h)	Trip Distance (km)
Mean	11:14	40.87	6,1	21,8
Std.Dev	3:20	134.28	1,3	9,6
Min	2:10	3.00	3,9	4,5
Q1	9:47	5.00	5,1	14,5
Median	10:10	6.00	5,9	20,5
Q3	13:18	7.00	7,0	29,5
Max	23:26	848.00	9,2	47,0

Table 52. Statistics of the group 5 (N=572) of trips of the Regensburg city car-sharing service

	Start time	Booked in advance (h)	Duration (h)	Trip Distance (km)
Mean	17:41	47.59	3,4	29,8
Std.Dev	3:52	153.30	1,1	7,8
Min	9:47	1.00	1,0	12,5
Q1	15:57	2.00	2,7	24,5
Median	17:02	3.00	3,4	28,5
Q3	19:43	4.00	4,2	35,5
Max	24:48	772.00	6,0	56,0

Table 53. Statistics of the group 6 (N=767) of trips of the Regensburg city car-sharing service

	Start time	Booked in advance (h)	Duration (h)	Trip Distance (km)
Mean	13:18	92.71	5,3	41,0
Std.Dev	4:51	299.91	1,8	14,6
Min	0:04	1.00	1,2	7,0
Q1	10:26	4.00	3,9	30,0
Median	13:50	5.00	5,1	41,0
Q3	16:09	7.00	6,8	51,0
Max	23:42	1998.00	9,1	72,0

Table 54. Statistics of the group 7 (N=556) of trips of the Regensburg city car-sharing service

	Start time	Booked in advance (h)	Duration (h)	Trip Distance (km)
Mean	14:49	86.29	2,7	14,3
Std.Dev	4:23	291.54	1,2	7,7
Min	0:28	1.00	0,8	4,5
Q1	10:02	1.00	1,8	8,5
Median	14:02	2.00	2,5	12,5
Q3	17:28	4.00	3,5	18,5
Max	24:51	1997.00	6,5	42,0

Table 55. Statistics of the group 8 (N=886) of trips of the Regensburg city car-sharing service

As far as the District's car-sharing service is concerned, we will also first analyse the trips made during weekdays and which correspond to the first 5 groups. In the first category, we have trips that are made in the morning, with a higher tendency towards early hours (half of the trips are made before 10:30 am). They have an average duration of about 9 hours and a distance of about 100km. Since there are two major cities at a similar distance (Nuremberg and Munich), and these are trips which are booked some time in advance, this group could be business trips to either of these two cities.

The second category contains journeys similar to the above but of a shorter distance, around 36 kilometres. The duration and the advance booking also seem to indicate that they may be associated with work activities, since, like the previous ones, they have an average duration of more than 9 hours.

The third category also corresponds to trips that take place in the morning, although at a slightly earlier time (half of them are made before 9:30). However, these are much shorter trips in both distance and time, with these averages standing at almost 3.7 hours and 22.9 kilometres, in that order. Since they seem to be more spontaneous trips (they are booked about 3 hours in advance (median)) their purpose can be related to activities such as shopping or personal issues.

The fourth cluster groups trips that take place in the early afternoon, but with significantly longer durations, distances and advance booking than in the previous case, with median values of 5.7 hours, 52km and 5 hours, respectively. With these characteristics, these trips could be associated mainly with leisure activities or shopping.

The last group related to weekdays is similar to the previous one, with the exception that they are trips made in the afternoon, mostly between 5 and 8 pm. The duration and distance are slightly lower, although the average time in advance of the booking is similar. Given these characteristics, they can be trips associated with activities such as leisure and shopping.

As for the clusters related to weekends, the first of these corresponds to longer and more lasting trips, with 70.1km and 8 hours on average, respectively, although with considerable variation, especially in the case of distance. The latter suggests that there is also a greater variety of destinations. They mainly start between 10 and 12 in the morning and are usually booked the night before. They thus seem to be more related to long-term leisure activities.

Finally, the second group of weekends corresponds to more spontaneous and shorter trips, which are booked about 4 hours in advance, have an average duration of 4 hours and an average distance of 31kms. These data show that they can be mainly leisure and shopping trips.

	Start time	Booked in advance (h)	Duration (h)	Trip Distance (km)
Mean	11:49	27.81	8,9	107,4
Std.Dev	4:31	66.32	2,5	19,0
Min	5:28	3.00	3,3	74,0
Q1	8:10	7.00	7,1	90,0
Median	10:33	9.00	9,1	108,5
Q3	13:09	11.00	11,1	122,5
Max	21:28	403.00	13,6	136,0

Table 56. Statistics of the group 1 (N=235) of trips of the Regensburg district car-sharing service

	Start time	Booked in advance (h)	Duration (h)	Trip Distance
Mean	9:22	17.70	9,9	36,5
Std.Dev	3:06	44.91	2,0	16,9
Min	4:28	6.00	6,7	7,5
Q1	8:42	8.00	8,2	23,5
Median	8:26	9.00	9,5	35,5
Q3	10:05	11.00	11,3	49,8
Max	22:45	331.00	14,0	77,0

Table 57. Statistics of the group 2 (N=95) of trips of the Regensburg district car-sharing service

	Start time	Booked in advance (h)	Duration (h)	Trip Distance
Mean	10:01	15.58	3,9	22,9
Std.Dev	2:59	56.23	1,4	10,0
Min	1:00	2.00	1,0	7,5
Q1	9:31	2.00	2,9	14,0
Median	9:29	3.00	3,7	22,3
Q3	12:46	5.00	4,9	28,5
Max	14:01	426.00	7,0	51,0

Table 58. Statistics of the group 3 (N=228) of trips of the Regensburg district car-sharing service

	Start time	Booked in advance (h)	Duration (h)	Trip Distance
Mean	14:03	19.99	5,5	54,7
Std.Dev	3:01	64.24	1,4	16,1
Min	6:56	2.00	2,1	26,5
Q1	12:22	4.00	4,7	40,0
Median	14:49	5.00	5,7	52,0
Q3	16:23	6.00	6,4	67,5
Max	23:15	444.00	9,1	101,5

Table 59. Statistics of the group 4 (N=153) of trips of the Regensburg district car-sharing service

	Start time	Booked in advance (h)	Duration (h)	Trip Distance
Mean	17:29	10.21	3,5	21,8
Std.Dev	2:18	45.12	1,1	11,8
Min	14:37	2.00	1,5	7,5
Q1	16:47	2.00	2,6	12,5
Median	17:13	3.00	3,3	18,0
Q3	19:58	4.00	4,2	28,0
Max	24:53	421.00	7,5	65,5

Table 60. Statistics of the group 5 (N=233) of trips of the Regensburg district car-sharing service

	Start time	Booked in advance (h)	Duration (h)	Trip Distance
Mean	11:56	26.17	8,3	71,3
Std.Dev	3:56	77.32	2,4	29,1
Min	5:45	3.00	3,3	9,0
Q1	9:56	6.00	6,3	53,5
Median	10:13	8.00	8,0	70,5
Q3	12:18	10.00	10,1	89,5
Max	22:37	427.00	14,0	136,0

Table 61. Statistics of the group 6 (N=29) of trips of the Regensburg district car-sharing service

	Start time	Booked in advance (h)	Duration (h)	Trip Distance
Mean	14:44	11.11	4,4	34,7
Std.Dev	4:51	42.27	1,7	18,8
Min	0:12	2.00	1,0	7,5
Q1	11:47	3.00	3,0	18,5
Median	13:26	4.00	4,3	31,5
Q3	16:20	5.00	5,7	47,5
Max	24:34	374.00	11,0	88,5

Table 62. Statistics of the group 7 (N=276) of trips of the Regensburg city car-sharing service

3.3.4. Analysis of the impact of car-sharing in private car ownership

In Regensburg, we also made an analysis similar to that presented in Section 3.2.4 to gain more insights into the importance of car-sharing in car-ownership in this city, as it is one of the research question posed for its test case. The dependent variable is again the number of cars in the household, and it has been divided into the same three categories: 0, 1 or 2 or more cars in the household. The number of individuals who fall into these categories is 335, 1432 and 729, respectively. Therefore the a priori probabilities of each would be 13.4%, 57.3% and 29.2%, in that order. The independent variables considered are the following: sex, age, labour status, education level, household income, household type, household size, number of motorbikes in the household, number of bikes in the household and use frequency of car-sharing service. The variable "City part" has not been included for the same reasons as those given in the Leuven case study. In addition, the variable type of public transport passes has also been eliminated because in preliminary analyses, we have found that it had a high impact on car-ownership. However, as we have seen previously, there were certain types of passes that had a high correlation with the use of car-sharing, so in an analogous way to "City Part", we have eliminated it to allow us to see the influence of the frequency of use of car-sharing services on car-ownership.

The construction of the decision tree model has been done in an analogous way, although in this case, the parameter cp was set to 0.012 and the parameter min_split to 50. The resulting decision tree can be seen in the next figure. The colour codes and the terminology are the same used for the Leuven Case Study.

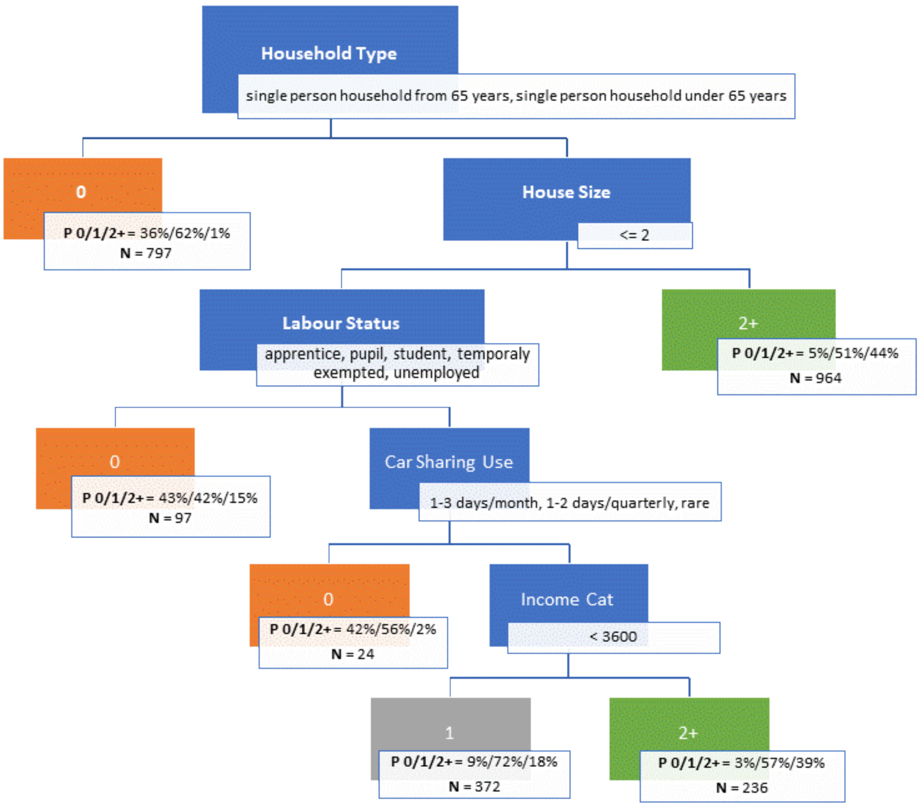


Figure 158. Decision tree for the estimation of car-ownership in Regensburg

In the tree, we can see that if a home is unipersonal, then its probability of having 0 cars rises to 37%. In the case of households with more than one person, the use of the car-sharing service has an influence on those households with working people, pensioners, and children. In that case, if the frequency of use of car-sharing services is Rare, 1-2 days quarterly or 1-2 days a month, then the probability of not having any cars in the household rises to 40%. Otherwise, the probability is 2% or 4%, depending on the household's income. If we look at the probability of having 2 or more cars, we see that if the individual is not in any of the car-sharing categories mentioned, then this rises from 10% to 18% or 40%, depending on the income level.

Although it may appear that car-sharing may not have an influence on households of more than two people, a similar analysis focusing only on households of this size shows similar results with car-sharing use as a relevant variable again. Specifically, it makes the probability of having zero cars go from 28% to 4%, or the probability of having two cars from 24% to 37% (use or non-use of car sharing, respectively).

3.4. Thessaloniki Case Study

3.4.1. The Dataset

In Thessaloniki's case study a large dataset of taxi routes collected by CERTH/HIT over the last 3 years is used for the better understanding of citizens' activities and their interaction with the current transport infrastructure of the city. Depending on which part of the transport system (passengers, taxis, bus, road networks, etc.) the data mining process is focused on, different results can be emerged. The combination of all the results from the different approaches form a more complete picture for the current transport conditions and generate insightful conclusions. For this purpose, the analysis process will start by defining the 3 different problems and analysing them separately. In the final step, the knowledge extracted through the analysis process will be used to develop useful tools for decision-makers on transport systems.

3.4.1.1. The dataset of OD matrix estimation.

The data used for obtaining the taxi OD matrix are the trips executed in November and December of 2019, including all hours of the day, and the focus of the analysis was between 8 and 9 o'clock. The longitude and latitude of the origin and the destination of each trip are matched to one of the 315 zones. If a trip is found to have started at zone i and ended at zone j, the OD for this specific hour at the matrix element OD(i,j) is incremented by one, that is OD(i,j)=OD(i,j)+1. This process is repeated for every trip. In this way, 24 hourly OD matrices for each day of these two months are created and therefore a total number of 60 OD matrices. Below, the actual OD element values of the OD matric for the interval 8-9 as well as their histogram are presented.

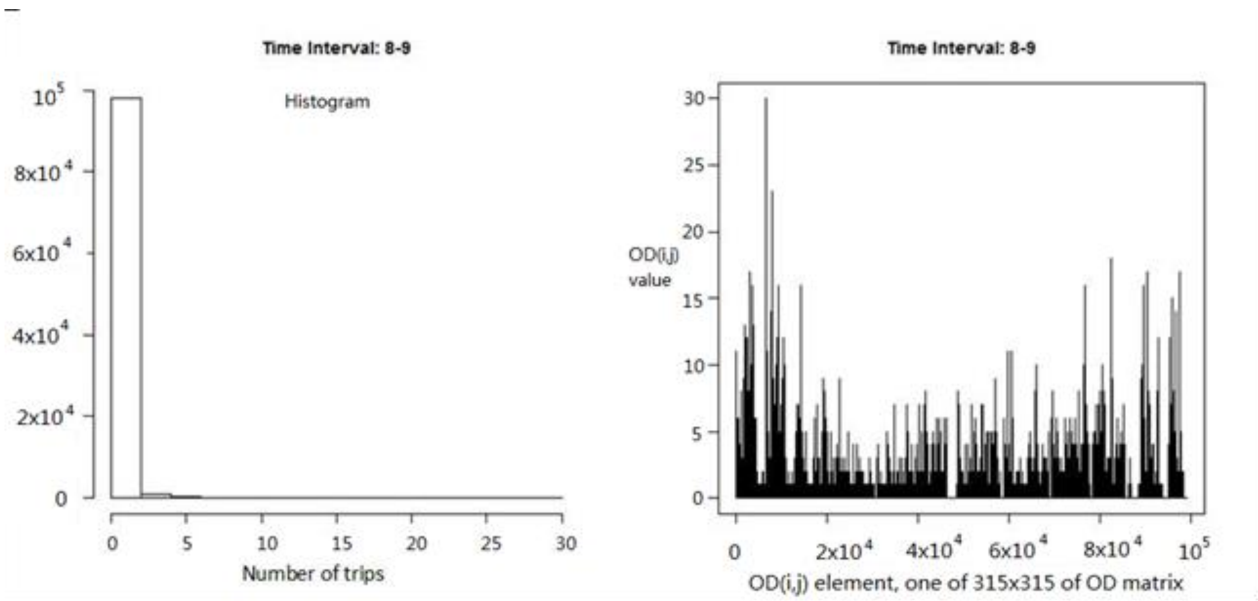


Figure 159. Values of the 8-9 OD matrix

It can be seen that the actual values of the hourly OD matrix for 8-9 o'clock time interval have a maximum of 30 and that their distribution seems considerably skewed as the largest percentage of the values are set to zero and only a significantly lower percentage is higher in value. The same fact can be observed in the 3D representation of the OD matrices for two random days.

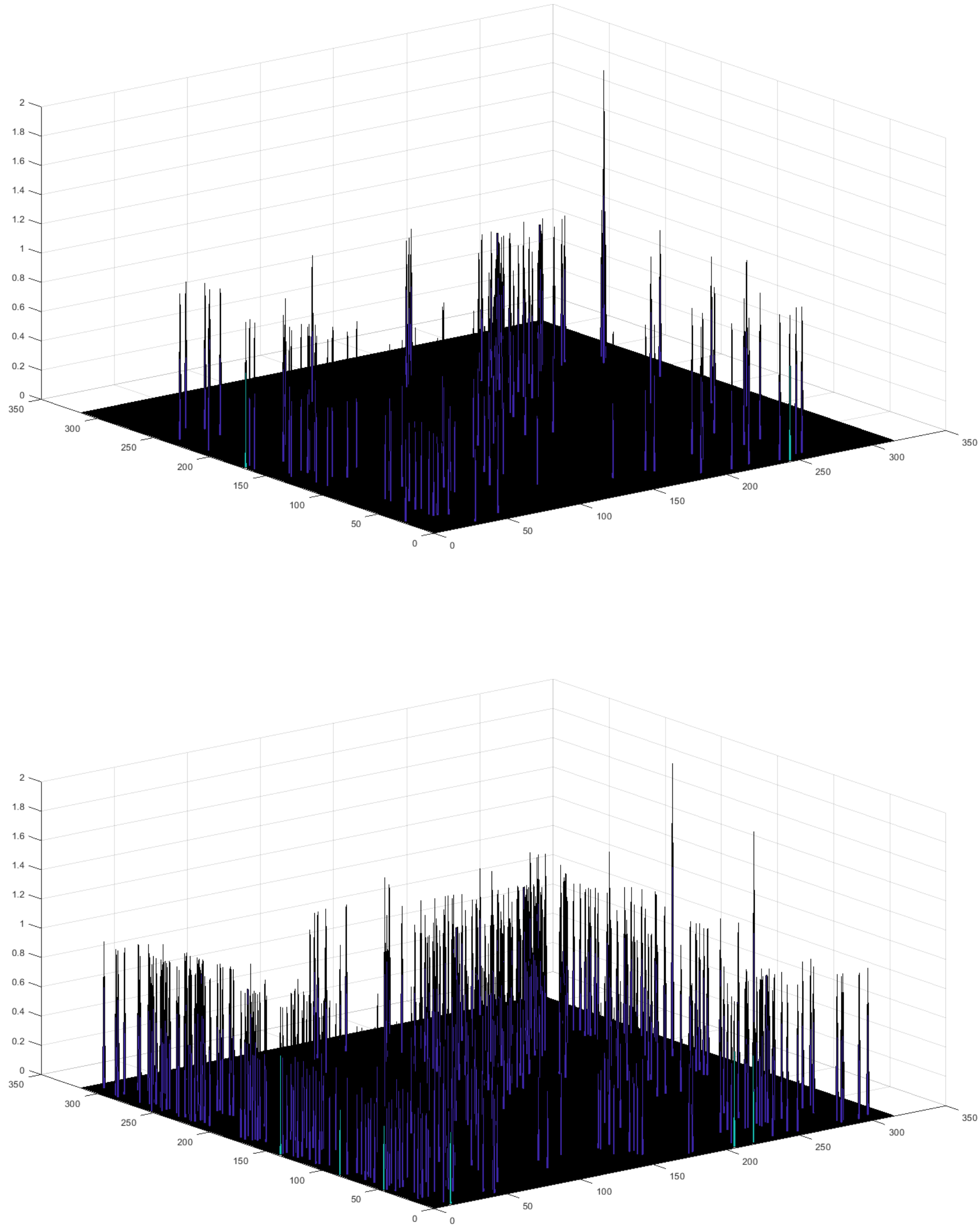


Figure 160. 3D representation of the OD matrices for the interval 8-9 for the days 11/11 (up) and 22/11 (down)

The same concentration of zeros is observed for every day in the dataset and for all hours.

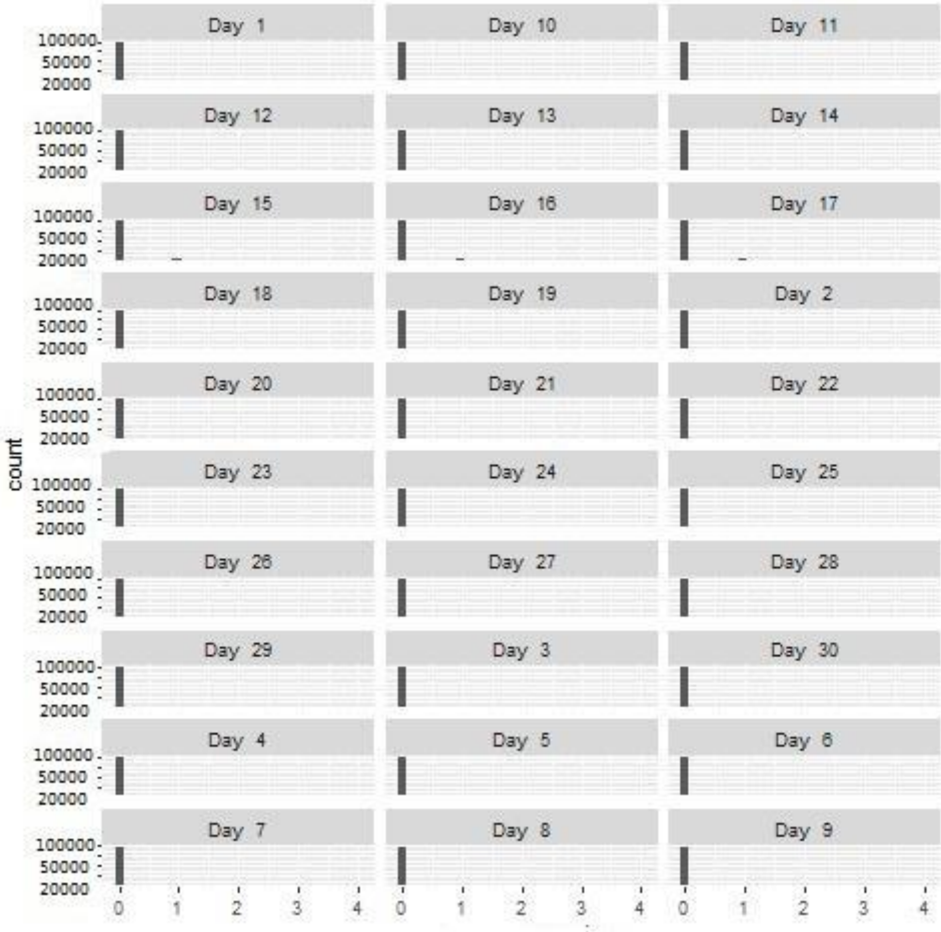


Figure 161. Distribution of taxi trips for the hour 8-9 and for every day during December

During the 60 days the number of trips ranged between 3405 and 6420, during the weekends and the weekdays respectively and with an average number of trip next to 5000 daily trips. The pattern is quite clear, excluding the weeks of December before and during Christmas, with an increase and decrease of trips respectively.

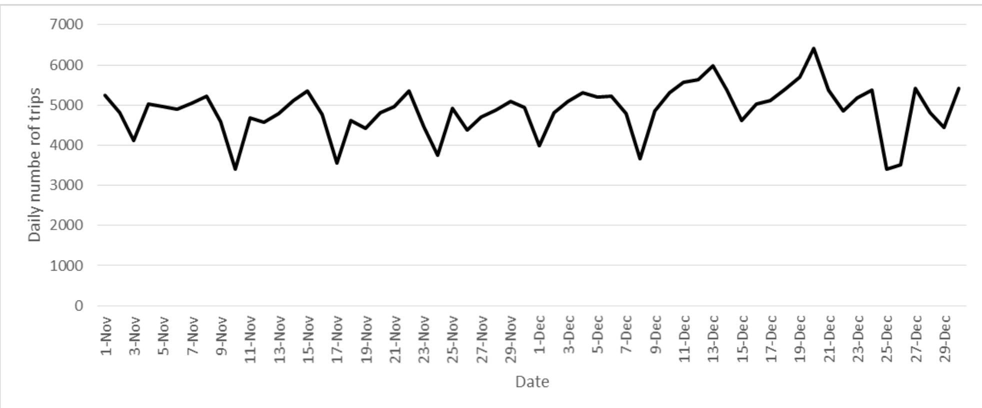


Figure 162. Taxi trips profile during November and December 2019 in Thessaloniki

With regards to the distribution of trips during the day there are two peaks, one in the morning (11-12) and one in the afternoon (17-18), while at 15-16 there is an off-peak period.

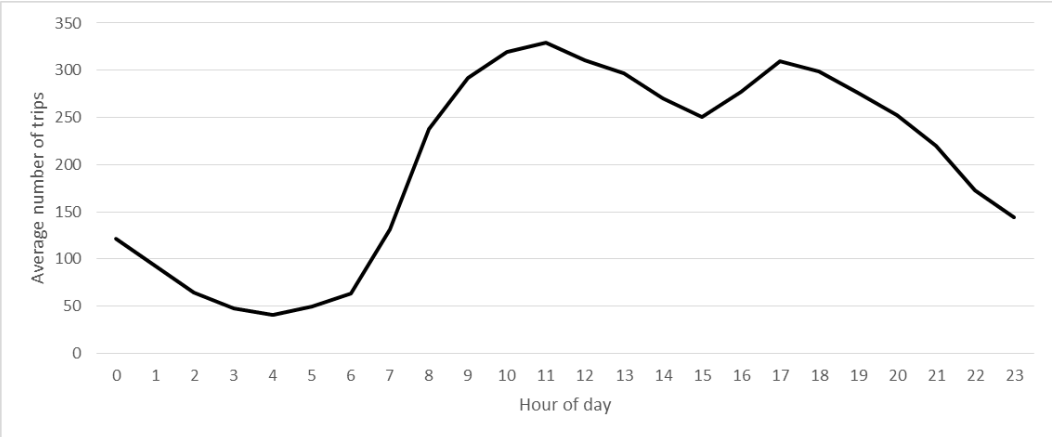


Figure 163. Daily taxi trips profile in Thessaloniki

3.4.1.2. The dataset for DRT and Ridesharing Area Analysis.

Taxi routes data that used in the case of DRT and ridesharing analysis were organized as presented in the following figure:

Index	start_geoc	tart_geog_	_start_zon	trip_start_timestamp	end_geog	nd_geog_	_end_zon	trip_end_timestamp	has_passe	uration_se	istance_m
0	40.6071	22.9532	50	2019-01-15 00:00:03.167	40.6363	22.9366	62	2019-01-15 00:06:42.007	1	399	3907
1	40.6385	22.9391	9	2019-01-15 00:00:16.000	40.6699	22.9437	11	2019-01-15 00:09:20.060	1	544	4067
2	40.6354	22.9541	66	2019-01-15 00:00:18.743	40.65	22.9457	26	2019-01-15 00:09:59.980	1	581	3151
3	40.6155	22.9754	49	2019-01-15 00:00:46.543	40.6355	22.9541	66	2019-01-15 00:35:15.557	0	2069	3868
4	40.6167	22.9586	7	2019-01-15 00:01:14.877	40.5932	22.963	65	2019-01-15 00:09:24.717	1	490	2771
5	40.6016	22.9731	4	2019-01-15 00:01:16.210	40.5991	22.96	65	2019-01-15 00:10:13.240	0	537	1566

Figure 164 : Dataset format for DRT and Ridesharing analysis

So, every row represents one observation-route and each column corresponds to one of the basic features collected in data pre-processing. The observation period of the current dataset which contains about 90,000 taxi routes is the 3rd week of January 2019. Therefore, this data came after elementary operation on primal raw FCD and contain a lot of outliers.

The basic features are:

- Starting point: Latitude, Longitude, and Timestamp.
- Ending point: Latitude, Longitude, and Timestamp.
- Total Distance Travelled during the route in meters.
- Total route time in seconds.
- Boolean that indicates if the route has a passenger or not.

Types of outlier patterns:

1. Close Origin and destination points (less than 500 meters) may be produced by wrong usage of GPS device from taxi driver.
2. Trip distances that are far higher that haversine or Euclidean distance.

3. Routes without passenger.
4. Routes with extremely large distance compared to trip duration.
5. Routes with extremely large trip duration compared to distance.

Thus, the basic step of the process is to clear those outliers and keep observations that are able to produce reliable results. Due to spatial nature of FCD, the Density-based spatial clustering of applications with noise (DBSCAN) algorithm is performed to achieve the filtering. DBSCAN is a clustering algorithm that produces non-linear spatial oriented clusters. Its most important feature is that it does not assign all the data points to a cluster. More specifically, it separates the points that belong to a cluster and those which are noise or outliers. Each point has three basic features, the trip distance, and origin and destination points. Using the origin and destination points the extra feature of Haversine distance for each point can be created. During the filtering process through the DBSCAN algorithm the primal trip distance and the Haversine distance were used as features.

Thessaloniki is the second largest city of Greece, and according to latest data the whole region has 1.1 million citizens. Furthermore, it is a city with a long tourist period due to many historical and commercial places. It is also the biggest college town in Greece with more than 100,000 students be a part of the social life. Besides that, central government policies target to initialize Thessaloniki as the main technology and commerce hub of Balkan Region. As a consequence, transportation system of the city is expected to be overloaded. However, it turns out that the operational capability of the city's transport system is a crucial factor (maybe necessary) to support the different sides of both social and business activity.

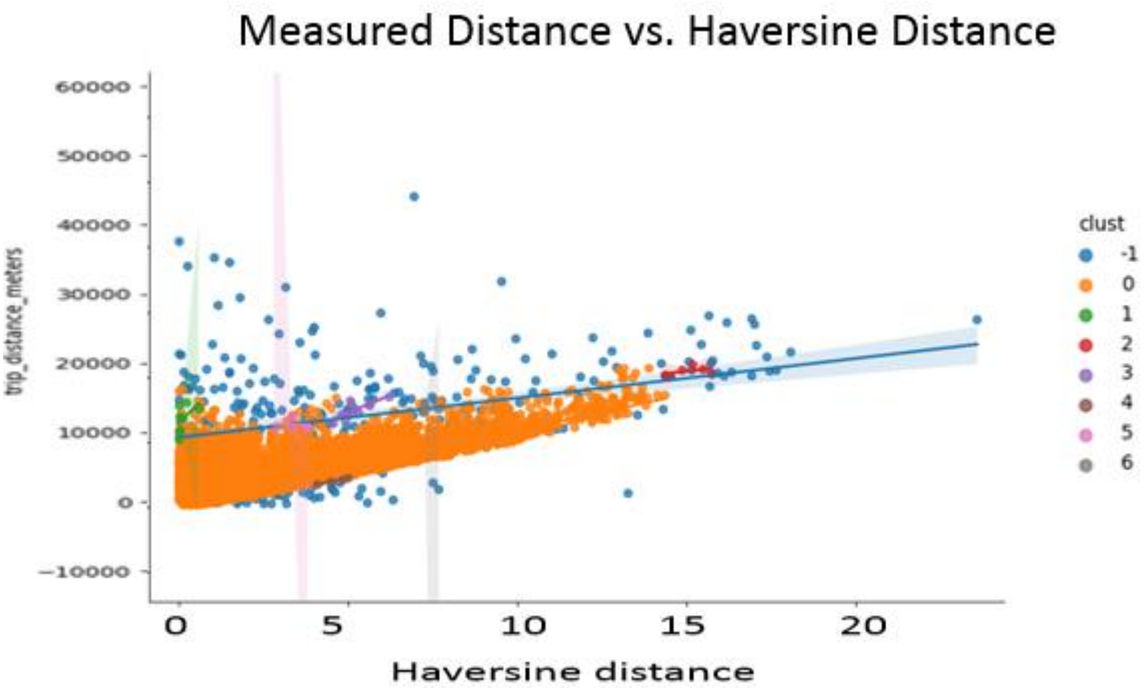


Figure 165. Haversine Distance vs. Measured Trip Distance. (Blue points belong in cluster -1 are the outliers)

All these different sources generate a remarkably sensitive and uncertain environment for transport planners. Ordinary tools that analyse the behaviour of citizens along the day are not capable enough to map a detailed picture of the transport system.

Consequently, new methods need to be explored for making possible to get more detailed insights about routing patterns across the city. Luckily, more and more cities around the world tend to understand the importance of data mining technologies due to their interest in progressive and novel solutions on how they plan the transport network. So, by taking advantage of the latest success which Data Science achieved, they aim to redefine the way they operate to fulfil people's needs.

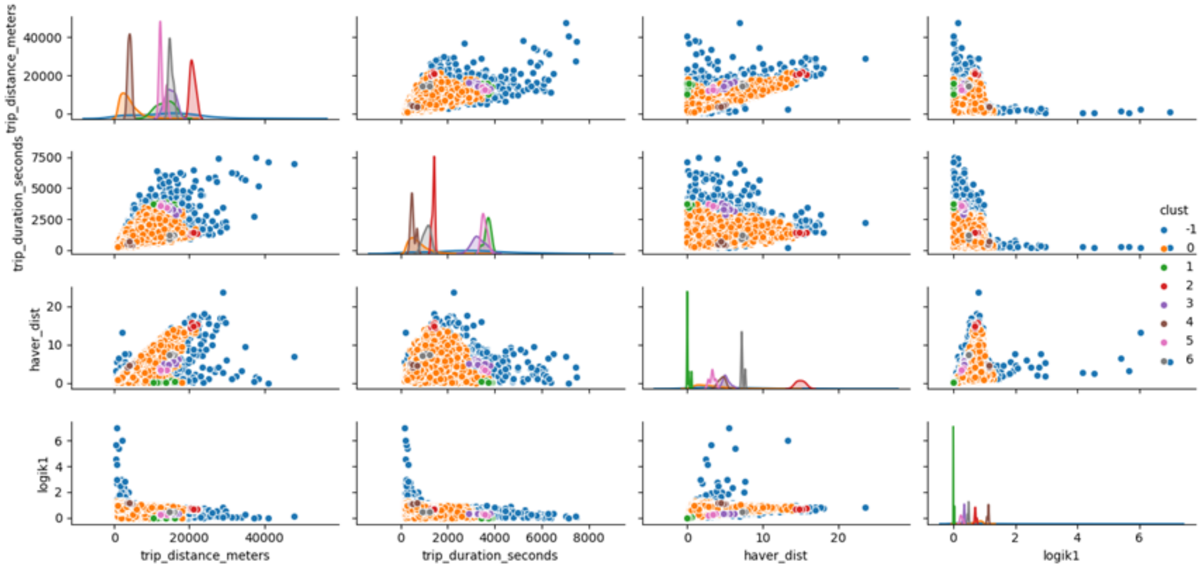


Figure 166. Paired Plot of basic features with outliers (Blue points) highlighted.

In Figure 166, the results emerged from the observation analysis are presented. Only12.800 trips out of the 90.000 can be considered as reliable observations. A closer look in that Figure 166 reveals that:

- DBSCAN efficiently detect outliers of Type-2 according to distance in meters and haversine distance pair. Blue points are those which do not follow that correlation of Orange points.
- Pair plot between distance in meters and duration shows that Type-4 and Type-5 outliers also removed.
- Logik1 variable is an extra feature that came after detailed and careful outliers' removal by hand, and based on the ratio of the trip duration divided by the trip distance and normalized with haversine distance. In fact, this measure is used to make the filtering of Type-4 and 5.
- 49.4% of the 90.000 trips were removed because they do not have passenger.
- The rest 50.6% were filtered based on the previous criteria. Thus, from the trips with a passenger, only 26.5 % is considered as a reliable sample for the analysis.

The current methodology explores the possible ways of usage for this type of data. In particular, based on the prior knowledge of network situation and a dataset of taxi routes, the methodology indicates methods through which the FCD data can contribute to the identification of some mobility patterns. Origin-Destination (OD) matrix is one of the most widely known methods for mobility patterns identification. It portrays, in a highly precise and informative way, the citizens' mobility behaviour like common trips that could be performed by ridesharing service. Data science part comes from the desire to get a more detailed representation of those matrices. By taking OD matrices on different periods of the day and rearranging the zones additional results that enrich and support the decision-making stages may be extracted. Other features like route trajectories are also considered as an important characteristic in order to create by taking into account the similarity of paths. Finally, the methodology describes a complete framework taking advantage of those trips and quantifying operational characteristics of a DRT system.

3.4.2. OD matrix estimation and comparison

The matrices obtained are presented below.

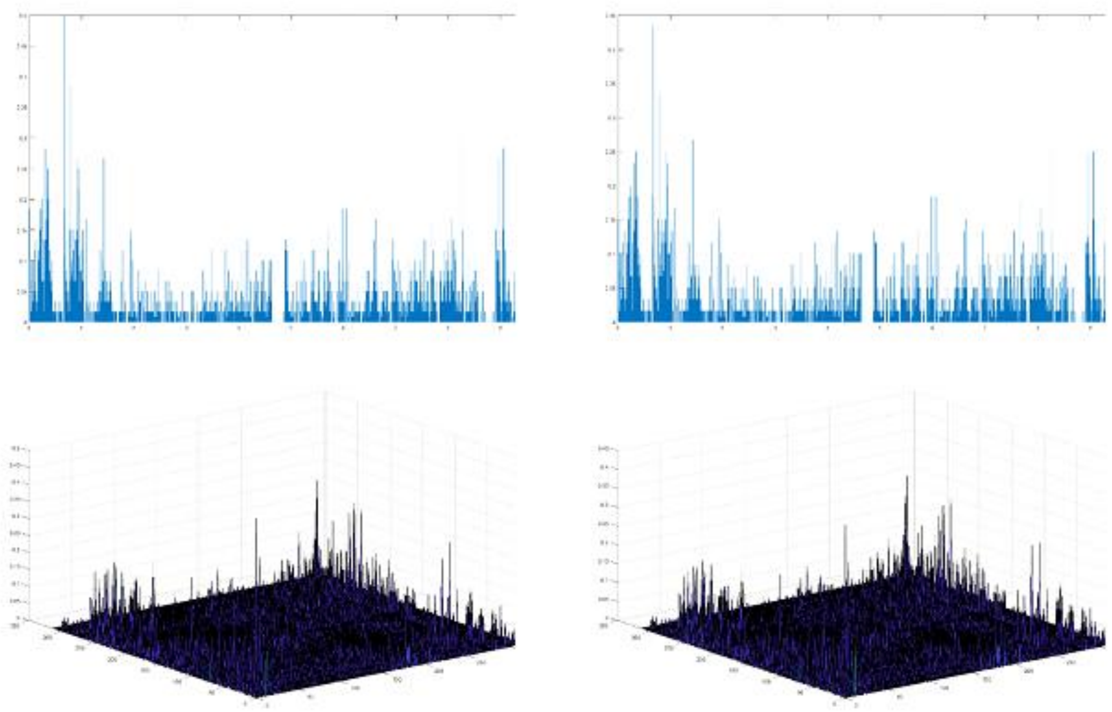


Figure 167. Representative 8-9 taxi OD matrices of Thessaloniki using the mean (left) and the e-med (right) methods

Since two representative OD matrices for the hour 8-9 for the two months mentioned are estimated, it is needed to compare these two matrices with each one of the original ones for each 8-9 hour of all of the 60 days of the two months to evaluate how close they are. The conditional entropy is the selected metric. In our case, each time we compare, either ε-med OD or the mean values OD will take the place of X and each of the real 8-9 ODs of each one of the 60 days of the months will be the Ys. Each conditional entropy referring to each 8-9 OD of every day is added and the two values are compared. Whichever value is smaller, may suggest that the corresponding representative OD is more suitable for the description of the traffic flow in the network of the 315 centroids. It seems that the sum of the conditional entropies of each 8-9 real OD matrix while X is equal to the new ε-med OD is smaller than the sum of the entropies of each of the 8-9 real OD matrices while X is equal to the mean values of the real 8-9 ODs. In the case of the ε-med values, the sum of the conditional entropies was estimated to be equal to 68.72 whereas in the case of the mean values, the sum of the conditional entropies was estimated to be equal to 69.19. This indicates smaller uncertainty for the real ODs if the ε-med OD is known which may indicate that it can represent the real 8-9 ODs more efficiently, with improvement of 0,67% in the case of the ε-med regarding the uncertainty of its values in reference to each one of the 60 real ODs, in comparison to the mean values case. As we can see, the conditional entropy can be used as an indicator of proximity between the representative OD and the real ODs and in our application, it has indicated that it can potentially be used as a measure to compare the two representative estimators.

In addition, other metrics for comparing the two obtained matrices, such as total number of trips, number of zeros and others are used. It can be observed that the values in the e-med matrix are slightly smaller than the ones in the mean matrix, since the distributions are asymmetrical with a high concentration of zeros.

	Mean	E-med
Total trips	237.6348	236.1514
Total number of zeros (%)	91.849%	91.849%
Average number of trips	0.002395	0.00238
Average number of trips (no zeros)	0.029381	0.029198
Maximum number of trips	0.5	0.43333

Table 63. Comparison metrics between the two obtained matrices

The same conclusions can be drafted when comparing the values of both matrices cell by cell.

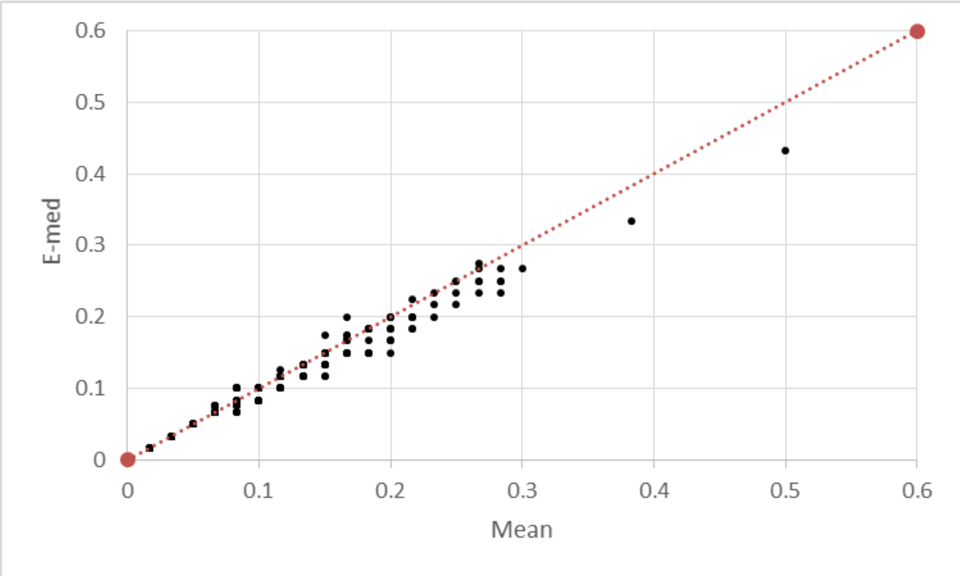


Figure 168. Comparison of values between the two obtained matrices

3.4.3. Case study of DRT and Ridesharing Area Analysis.

To define which areas are more suitable for DRT or Ridesharing lines it needs to combine both the KPIs defined above and some insightful visualizations about how demand progress through day hours. In Figure 169 interactive maps are presented that show the demand of the selected stops, for a specific day and peak hour. It shows a map divided into two parts; left one with the origin and right with destinations. Both contain an hour control bar and a popup charts in each district with the distribution of distance travelled. Finally, some of the bus lines are plotted in the maps to give intuition about the reasons that passenger prefer taxi over bus and the overall spatial – temporal distribution of demand.

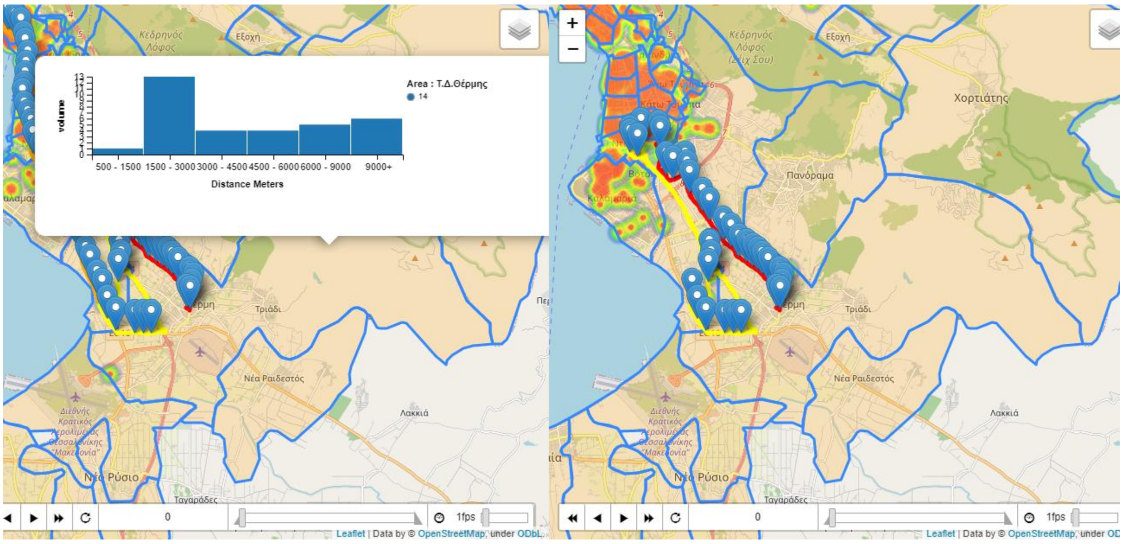


Figure 169. Demand and Destination on several districts of Thessaloniki for Taxi trips and distribution of trip distance (Control bar with Day Hours included).

The next step of the validation procedure is to apply and examine the proposed methodology, in a smaller case scenario. In this scenario, two regions of Thessaloniki were selected in order to investigate if in those areas these areas are suitable for the implementation of a DRT system.

3.4.3.1. Geographical scope

The first region is Thermi (A, exurb), which is a peri-urban area, sparsely populated with increased business activity. On the other hand, region of Kalamaria (B, suburb) is described as a densely populated suburban area of Thessaloniki, with increased household activities and close to city's ring road.

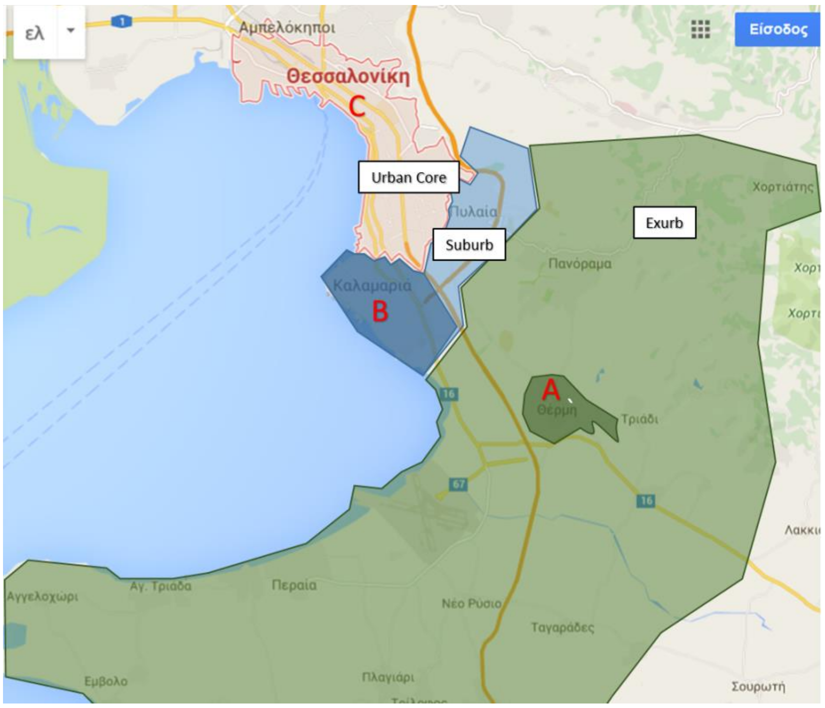


Figure 170. Areas of the Thessaloniki pilot

More specifically, the Municipality of Thermi is laid 15 km from the city centre of Thessaloniki on an area of 385.3 km2 from the south-eastern coast of Thermaikos gulf to the mountains of Hortiatis and Holomontas. The total population of its three municipal units (Thermi, Mikra and Vasilika) is 53.070 (2011 census. As it is located close to the airport, in the eastern side of the city's urban area, the Municipality of Thermi serves as a commuting point of 1,3 million visitors (year 2013). Its modal split is 72% for cars, 8% for motorcycles and only 20% for public transport.,

The Municipality of Kalamaria is one of the most populated municipalities not only in the metropolitan area of Thessaloniki but also in Greece. Although it is ranked third in Thessaloniki's greater Metropolitan area with 91,279 inhabitants (2011 census, its real population is estimated of about 120,000 inhabitants. It is located on a coast area of about 7,2 km2, 7 km from the city centre at the south east side of Thessaloniki. It borders with Pilea-Hortiatis on the southeast and with Thessaloniki on the north while its northwest side is surrounded by sea. The municipality consists of 11 districts, namely: Kifisia, Karampournaki, Kouri-Katirli, Aretsou, NeaKrini, Ag.Ioannis, Bizantio, Ag. Pandeleimonas, Navarxos Votsis, Foinikas, Derkon. Its modal split is 37% cars, 50% public transport, 7% motorcycles, 1% bicycle and 4% walking.

The trips between these two municipalities and the exurb, suburb and urban core areas of Thessaloniki are presented below (obtained from the transport model of CERTH-HIT).

Municipality	External	Exurb	Suburb	Urban core (C)	Total
Thermi(A)	206	1.786	2.461	**1.504**	5.957
Kalamaria(B)	204	800	7.295	**3.021**	11.320
Total	**410**	**2.586**	**9.756**	**4.525**	**17.277**

Table 64. Trips between the areas of the pilot

The modal split of the trips between the two municipalities and the urban core area of Thessaloniki is presented below, depicting in part the difference in the level of PuT service of each region.

	Thermi	Kalamaria
Bicycle	0%	1%
Walk	0%	4%
Codriver/Lorry/Motorcycle	8%	7%
Car Trips Core	72%	37%
Bus Trips core	20%	50%

Table 65. Modal split in the areas of the pilot

The good level of the public transport network of Kalamaria facilitates the trips of its inhabitants to the city centre. The public transport networks of the other areas are less dense. In Thermi there is only one bus that connects the municipality to the city centre, after one transfer in a peripheral hub. So, the use of public transport in Thermi is quite lower. As these Municipalities are located out of the urban core, their largest part can be considered as suburban. However, some of their districts belong to the exurb area of the city centre of Thessaloniki. Within the suburban area, there is limited possibility to reach the city centre without using the private car, especially at night, because public transportation does not operate after midnight. The predominance of the private car observed in these areas explains why the ride sharing and DRT services seem to be proper solutions.

3.4.3.2. KPIs calculation.

KPI name	Kalamaria	Thermi
Nts	3 (Car, bus, taxi)	3 (Car, bus, taxi)
Nbl	5 (5, 6, 3K, 1N, 2)	5(66, 36, 3K, 2K, 45)
Nbs/ km^2	20.6 stops/km^2	4.27 stops/km^2
IT	16 min/arrival *bus	20 min/arrival *bus (small busses)
Npa	According to Figure 3-4	According to Figure 3-4
Npla	According to Figure 3-4	According to Figure 3-4
LU	Household, shopping, leisure and entertainment.	Industrial- Business , shopping, entertainment
Other statistics	• 55% of trips start from Kalamaria end up in city center. • 38% of trips take place 7:00-12:00 (5 hours 20% of day) • 23% 1.5-4 klm trip distance • Over 42% trips over 5klm. • Almost all passengers preferred taxi was less than 300 meters from bus stop.	• 55% over 3klm • 58% of trips start from Thermi end up in city center. • Difficult road network for bike systems. • 48% of trips take place 7:00 – 11:00 (16% of day) • Almost all passengers preferred taxi was less than 200 meters from bus stop.

Table 66. Comparison among urban and peri-urban area.

Some conclusions about those results are highlighted below:

- Both regions of Kalamaria and Thermi have the same number of available options of public or private transport infrastructure. However, in Kalamaria it is more difficult to find parking as a densely populated region. Although Thermi is a more extended area, it is served by an equivalent number of bus lines. The same thing stands also for the number of bus stops and interarrival time.
- Both those areas have a significant part of routes that end up in municipality of Thessaloniki (city centre). Additionally, most of those routes performed during morning hours.
- It seems that Kalamaria is more suitable for ridesharing trips, micro mobility and bus lines. Micro mobility could take a share from that 23% of trips with less than 4km distance, and especially those who end up in city centre where bike network infrastructure is more mature. Additionally, due to the increased population and mobility activity in Kalamaria, the area seems to be more suitable for a ride-sharing service. Thermi needs a quite different approach on planning and design of such system. It seems more suitable for a DRT line due to the sparse demand density, same destination and poor public infrastructure Moreover, the fact that most of passengers don't prefer bus line, even when they are closer to 200 meters, reveal a systematic failure of the existing bus system to efficiently serve the area's inhabitants offering reliable and high service level.

Based on the characteristics of the two areas, the implementation of a DRT system will be analysed in Thermi, while a ride-sharing could be potentially implemented in Kalamaria.

3.4.3.3. DRT service in Thermi

As part of WP3's actions, the implementation of a DRT system was investigated for the case study of Thessaloniki. Based on the methodology framework analysed in the previous chapter, it is important to define the routes available for the DRT system.

In order to analyse the implementation of a DRT system in Thessaloniki, the number of applicable bus stops were examined. In those two experiments the number of stops extracted by the facility location step are 12 and 14. In Figure 171 the distribution of distance of each route is presented. The philosophy of a DRT service is not to stop in every station, but to stop only where users expressed their intention to use the system.

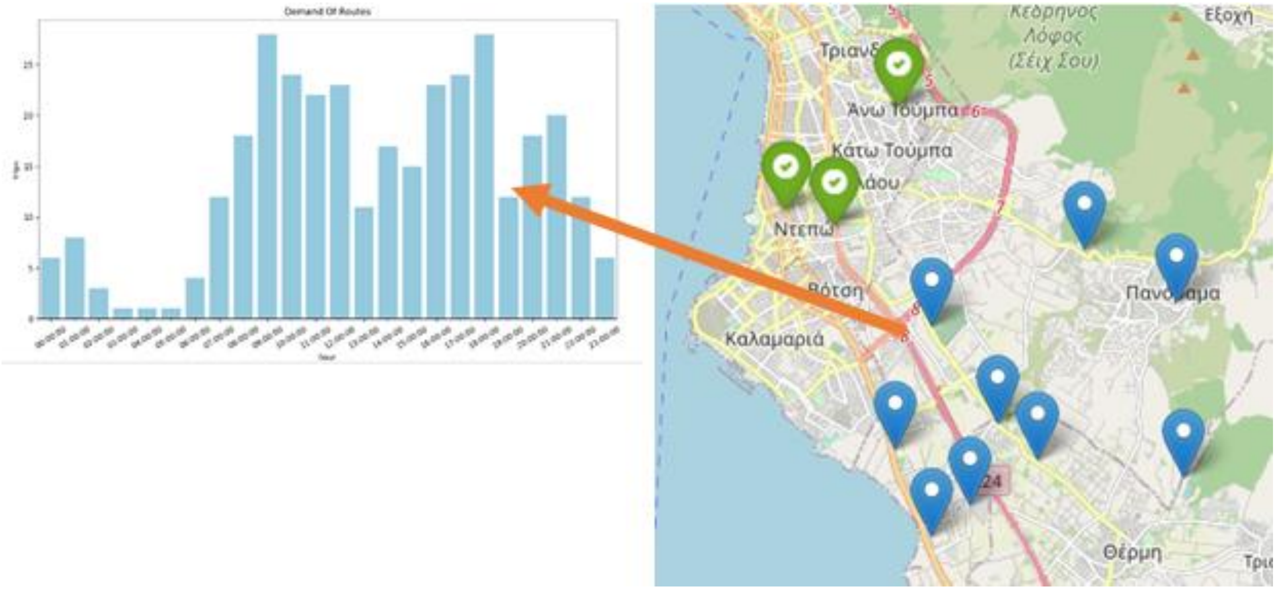


Figure 171. Sample Set of Facility Locations. (Blue points are the Pickup Stops and Green points are the pre-existing bus stops integrated to DRT line)

The discrete event simulation followed in the case study, helps to derive the distribution of trip distance, making the planning and scheduling steps of the procedure easier to be determined.

The other chart shows the frequency of number of stations that bus visited on each route. Figure 172, Figure 173 are corresponding to the 12 stops experiment and Figure 174, Figure 175 correspond to the 14 stops experiment. For example, by increasing the number of stops from 12 to 14 and involve more areas in the procedure, mean distance travelled increased from 10.8 km to 28.8 km. The fact that 57% of Thermi trips end up in municipality of Thessaloniki lead to the development of a DRT line with well supported bus lines which will serve the related area. Those results are valuable knowledge for decision-makers and planners, to create an integrated sustainable mobility system by reducing operational cost and maximizing fleet utilization. So, with the use of that framework, it's possible to find the parameters that optimize DRT system operation by considering all trade-offs involved. Additionally, increasing the number of stations, the uncertainty of trip distance (and total time) also increases. Therefore, it would be helpful to extend the PDTSP solved here to the PD-VRP. That would lead to a more general solution with better overall performance. The main idea is to break the decision process into smaller parts (k-means, facility location, TSP/VRP, simulation) and get a good estimation of the true operational cost and characteristics. In the current study, only FCD of taxi is available. Due to the small demand for taxi trips in Thermi, it is found that the total taxi demand can be satisfied with one DRT line. Of course, that will change if overcrowded bus lines and unsatisfied public bus passengers are considered. In case, of 14 stations that cover a bigger part of Thermi the average operational cost increases significantly as the uncertainty of trip duration gets higher. In fact, the number of stations is increased by 16%, the number of passengers that will be served is increased by 18%, and the total operational costs are expected to almost be doubled.

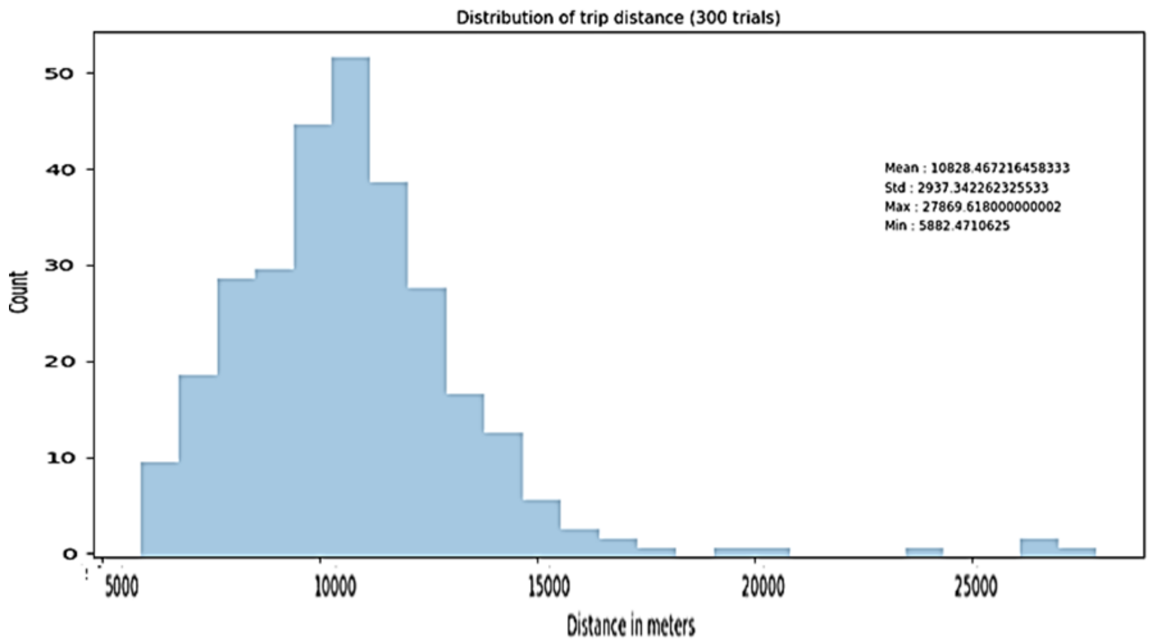


Figure 172. Distribution of trip distance for 12 DRT stops

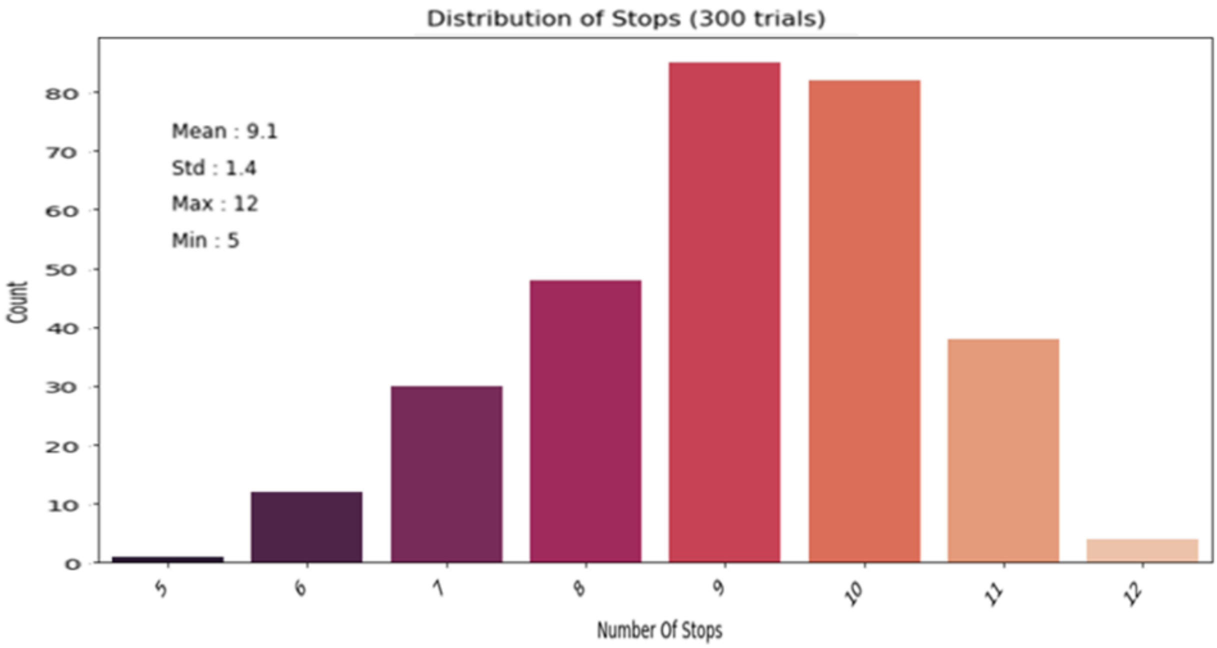


Figure 173. Distribution of Stops 12 DRT stops

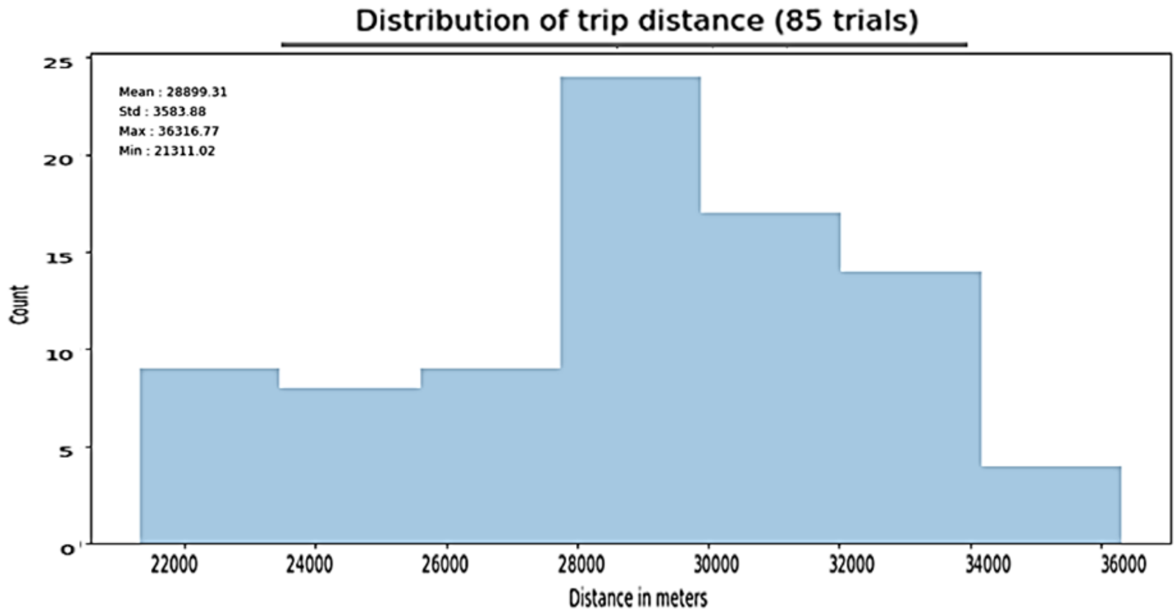


Figure 174. Distribution of trip distance 14 DRT stops

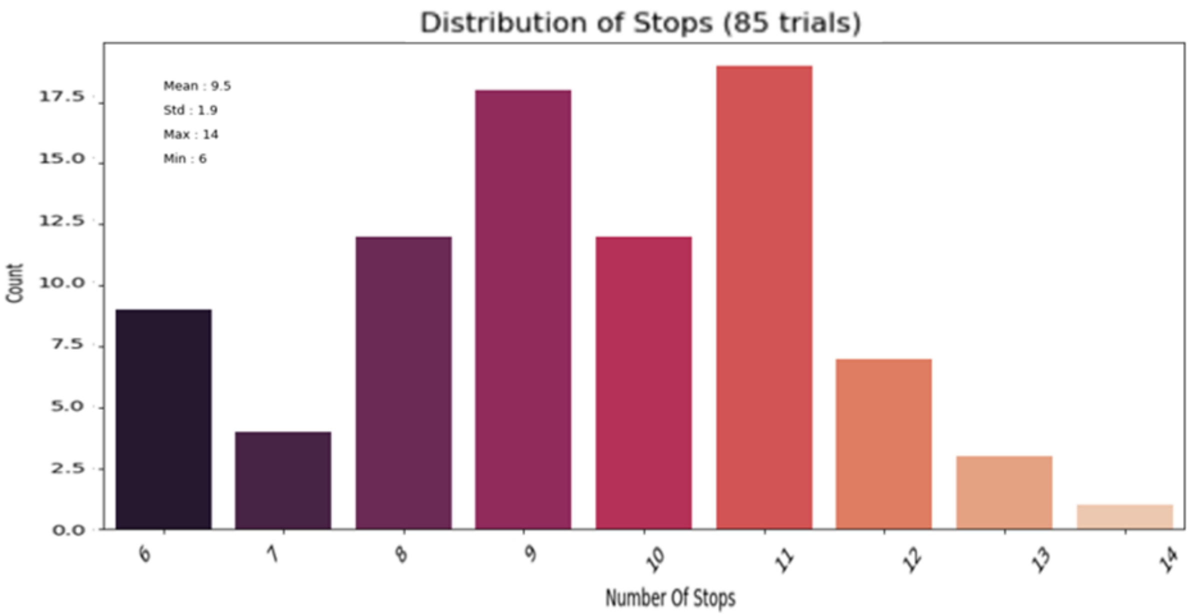


Figure 175. Distribution of stops for 14 DRT stops.

The distribution of distance can help to derive the average trip duration. Considering an average speed of 30km/h, it is easy to express the distance in terms of time (trip duration). Then we can assume that at one stop the arrival distribution of a passenger among two fixed buses arrivals is uniform. So, the probability of a passenger to take a DRT line rather than the bus line can be derived by subtracting the distribution of passenger arrival by the distribution of trip duration. To get an estimate of how much busses need to operate in a single line it needs to estimate that new distribution. Making that subtraction and simulating for all stops we take an average among all stops. The results presented on Figure 176. Those results refer to 12 stops DRT system because it was more economically efficient. By using two vehicles, the DRT line can serve over 62% of passengers earlier and more reliable than bus line. To get a greater share of 80% or more it needs more than 4-5 vehicles. So Thermi seems that need a DRT line with the stops of Figure 171 and 2 vehicles with a capacity of 20 passengers each.

That solution overcomes most of the problems mentioned in Section 2.2. Firstly, that line came as a supplementary on existing PuT network. So, avoids possible conflicts and further unexpected restrictions in the future. Furthermore, the current results of Figure 176 assume that the DRT line does not disturb the existing fixed bus lines schedules, by making the arrival between 2 arrivals of it. The most important part is that the DRT line is planned with respect to the demand for that area while the possible impact of demand uncertainties in operational characteristics are also considered. Consequently, the two most important reasons that lead DRT systems to failure (demand, and integration with existing PuT) are used as the backbone of the current planning procedure. Those results show how multiple data sources can be valuable for estimation of real demand along with the knowledge of schedules of current PuT infrastructure. In conclusion, the more data are available for demand, and the more integrated is the DRT line with existing PuT the lower is the probability of failure.

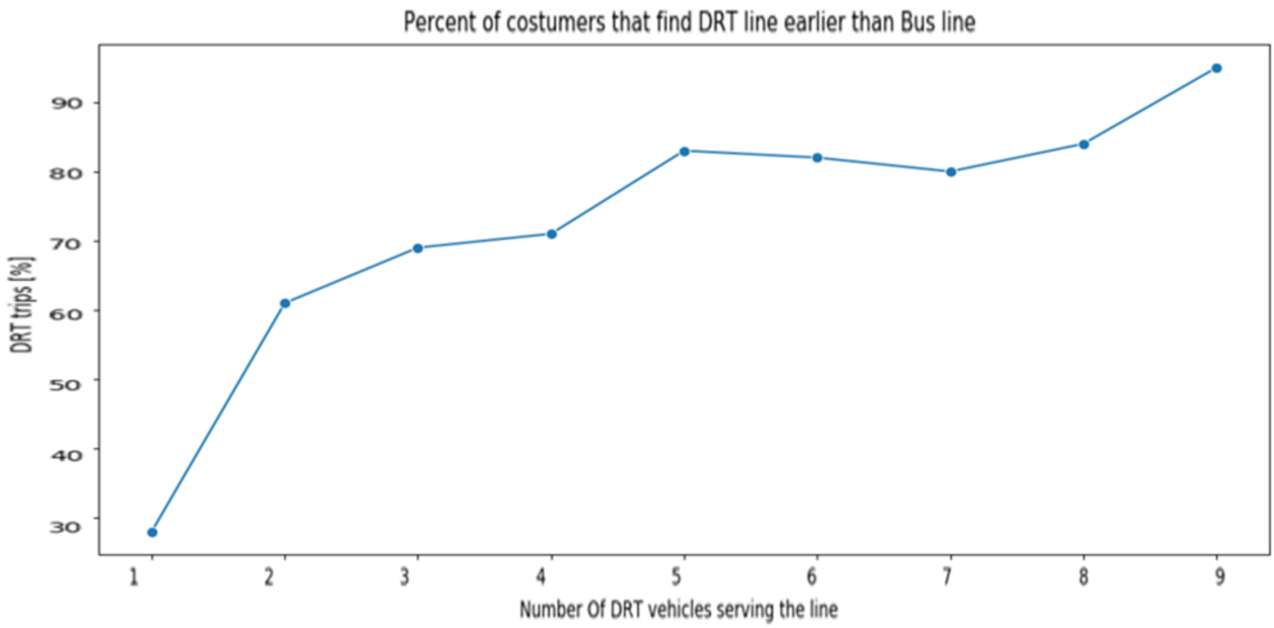


Figure 176. Percent of costumers that find DRT line earlier than Bus line.

3.4.3.4. Ride-sharing service in Kalamaria and other nearby municipalities

According to our methodology two trips with strong similarity would probably belong in the same cluster. That cluster indicate one possible ridesharing line or more. That method could provide planning steps of transport systems with insightful results and probability distributions about demand. It could support decisions like an input in any operational model as optimization programs or discrete event simulation. In following examples, we try to give some intuition about areas that has the potential for a mobility shift to ridesharing or micro mobility, changing the current transport situation.

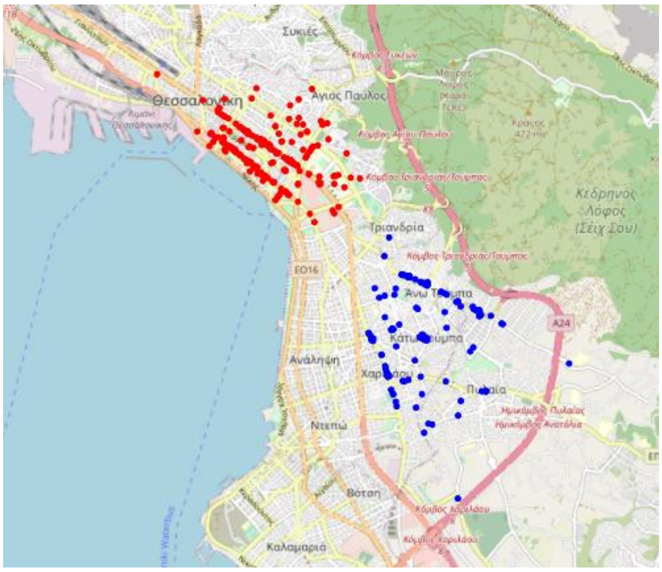


Figure 177. Sample cluster (Blue are the origin points, Red the destination)

The cluster presented in Figure 177 is a potential ridesharing cluster. That cluster starts from a region called Toumpa. Toumpa is a densely populated area with poor public transport or bicycle network infrastructure and thus is expected to be a potential area for the implementation of a ridesharing service. Furthermore, it has a lack of parking space. These factors reveal that this area is an appropriate district for public authorities to enforce ridesharing services.

According to the Figure 178 & Figure 179 during 7:00 – 19:00 the demand for taxi is nearly stabilized, with a small pick near 17:00. That probably comes as a consequence of many businesses and services that close near 17:00 – 18:00 and the corresponding congestion events in public transport systems. Thus, it seems that ridesharing services should orient their functionality to the business sector. Finally, on Sundays it does not seem that Toumpa has a significant demand due to the football games that take place in that area.

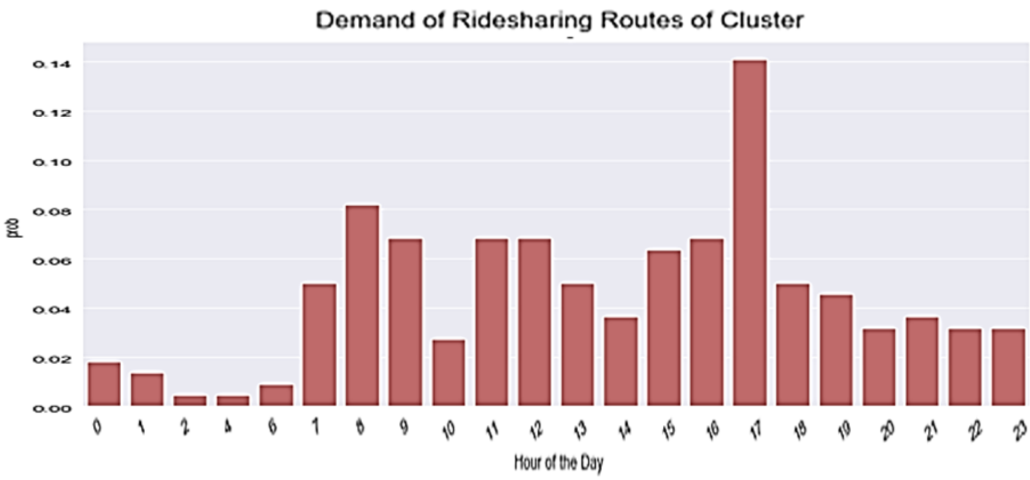


Figure 178. Distribution of Demand Over Time.

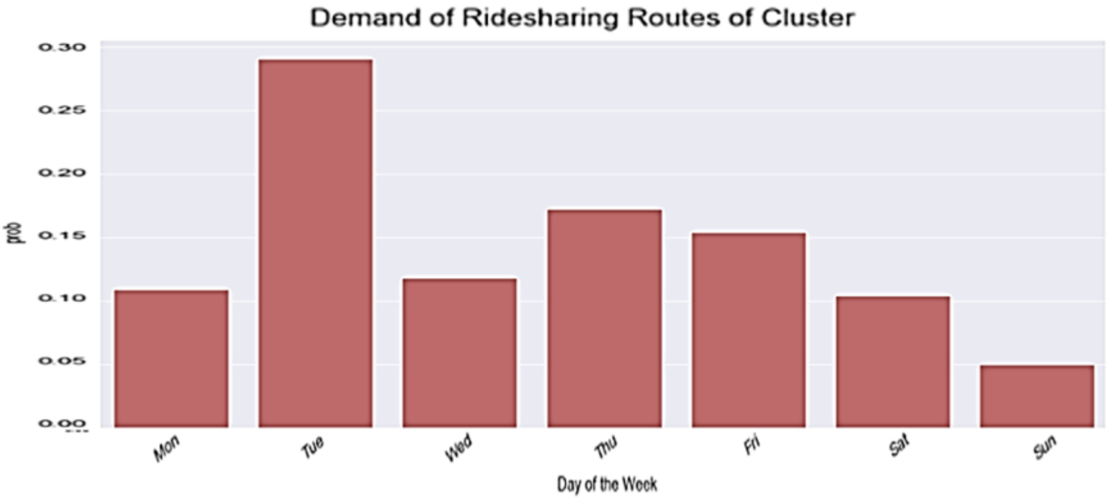


Figure 179. Distribution of Demand over Week

Trip distance is also an additional interesting insight. As shown in Figure 180 & Figure 181 mean distance of that trip is near 4 km. Due to poor bicycle network even routes near to 3km cannot be performed in that mode. The walking alternative even in 3 km case (~30min) is also not always possible or desirable. Those are additional factors that reinforce the conviction that this area is suitable for ridesharing service.

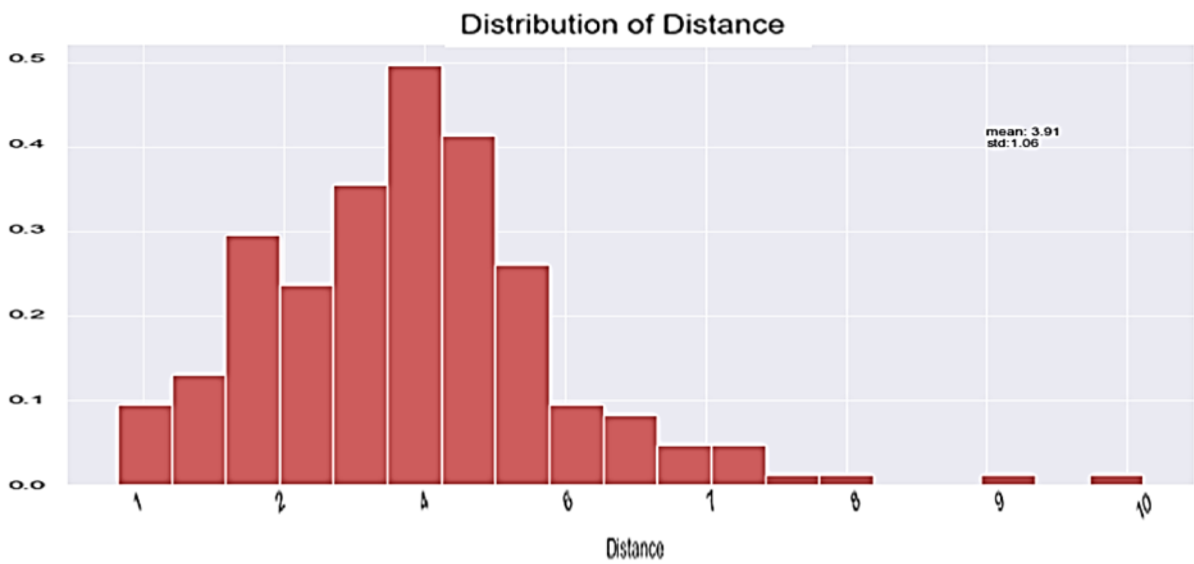


Figure 180. Distribution of distance.

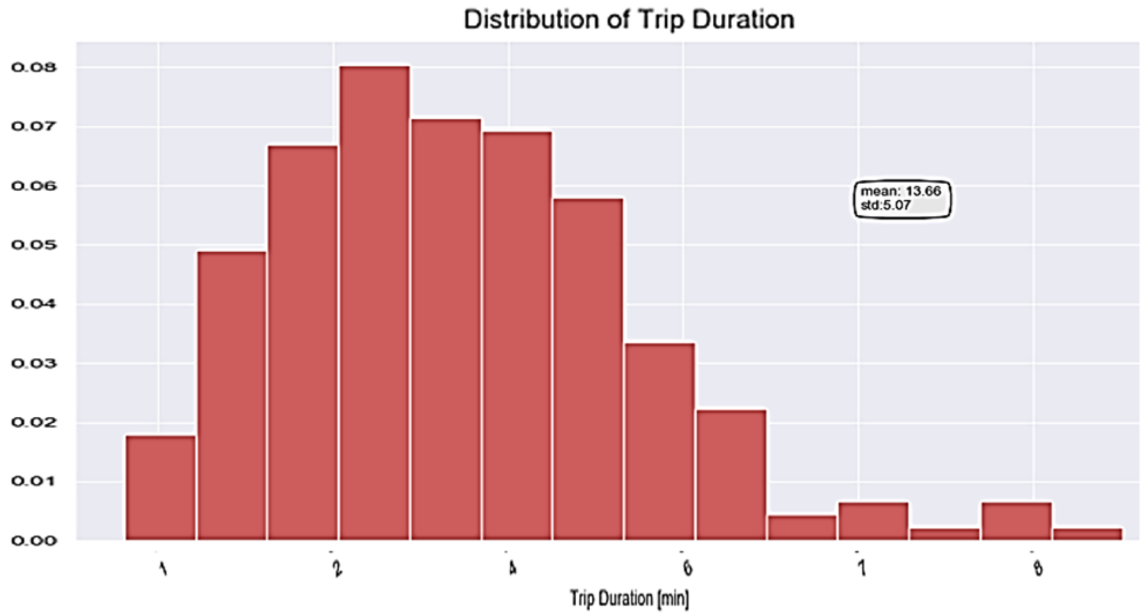


Figure 181. Distribution of trip duration.

Those data for demand distribution can be considered as a powerful input in the planning step of ridesharing systems. That system could take advantage of that knowledge with the use of appropriate algorithms and discrete event simulation tools. Consequently, it is important to use similar methods and support operational planning procedures with accurate inputs.

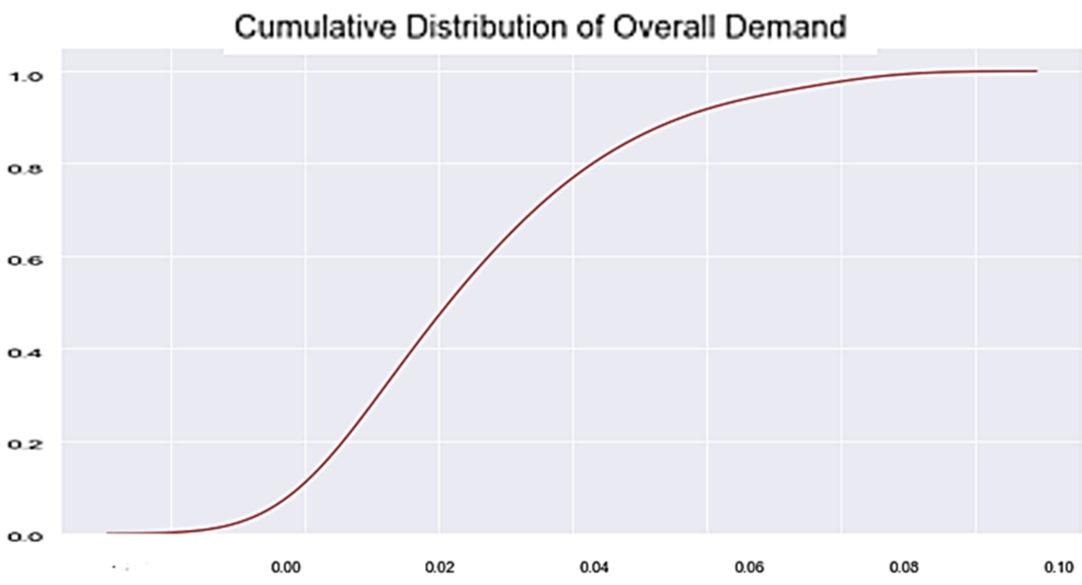


Figure 182. Cumulative Distribution of Cluster's Demand.

Besides of the results of the previous cluster example related to statistics derived, some additional interesting results that reveal the power of trajectory similarity measure are also presented. As shown in Figure 175 that method is able to summarize near 54% of the demand into 25% of the clusters. It forces in some way to give a more compact and reasonable form in that basic and rich feature that called path.

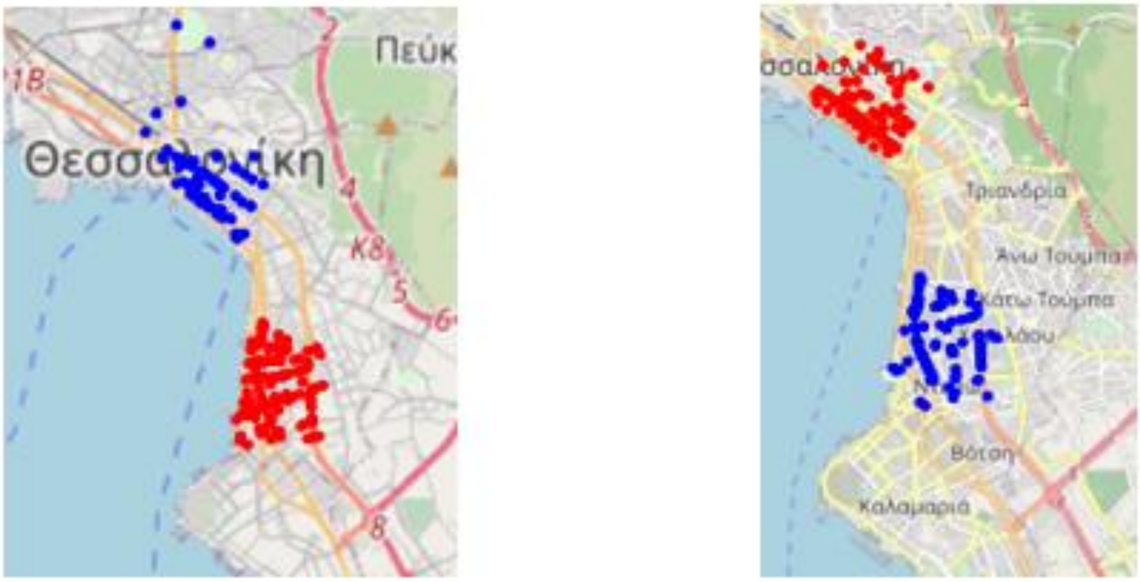


Figure 183. Mirror Clusters.

Interestingly, Figure 183 Left and right cluster are both in that top 25% of clusters which contain 54% of demand. It is important to mention that blue nodes depict the origin points while the red ones the destinations. More specifically, the Left and Right cluster of Figure 183 are the top 3rd and 4th clusters respectively. It is clear that the origin points from the right cluster are close to the destination points in the left cluster destination and vice versa. Additionally, that pair of clusters face an increased mean trip distance along with small standard deviation compared to other 'high demand' groups (Figure 184).

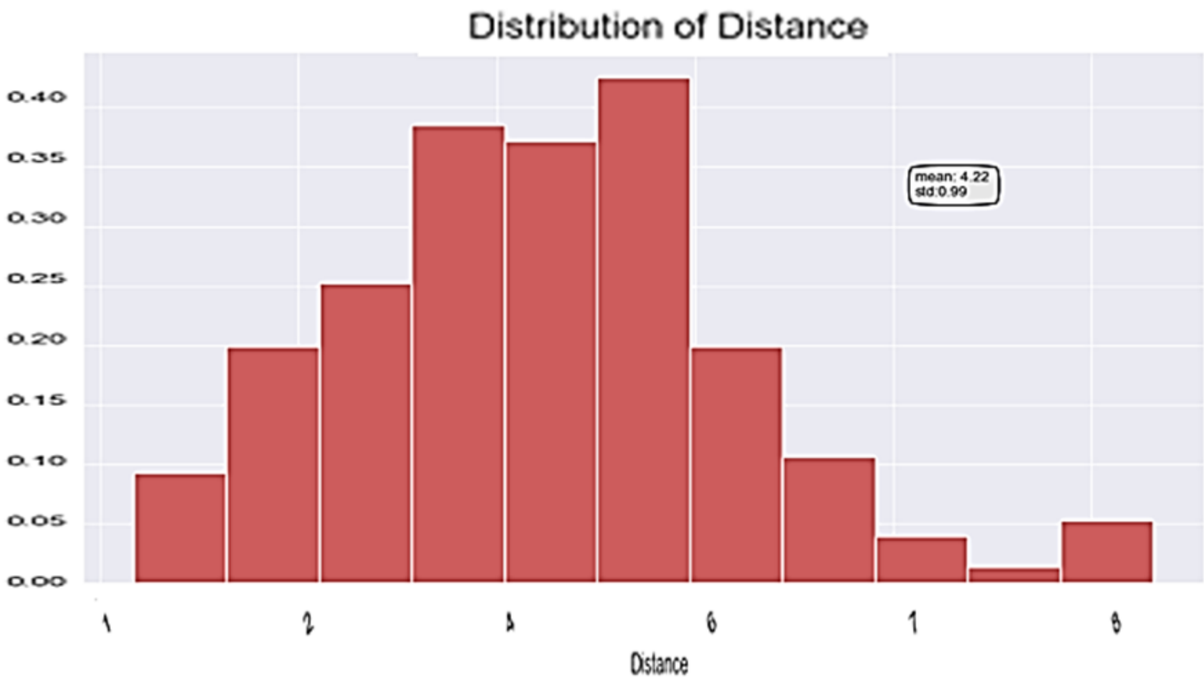


Figure 184. Distribution of distance.

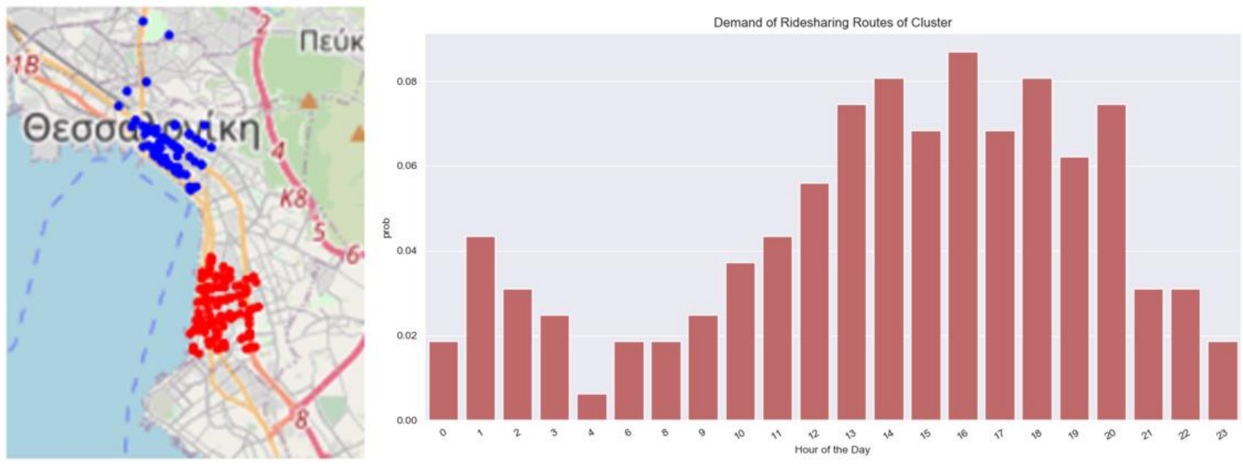


Figure 185. Demand of Left

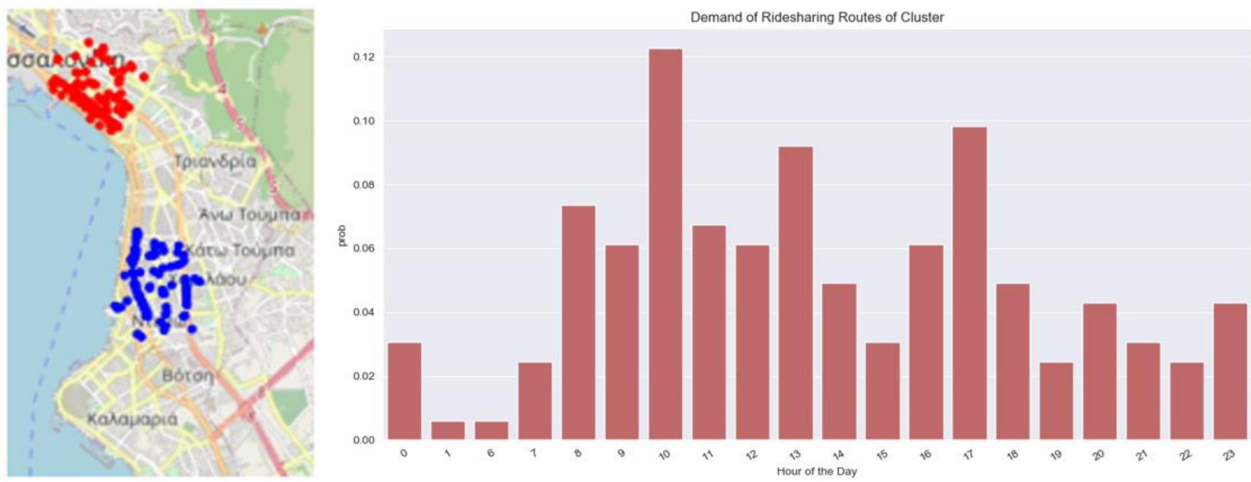


Figure 186. Demand of Right

Another interesting insight deriving from the investigation of this route in Thessaloniki, is the consistency of the trips. As it is depicted in the Figure 179 and Figure 180, the probability of users traveling from the one area to the other one is increased and fluctuates between morning and evening peaks times. Thus, it is concluded that produced trips from the one district is close to the attracted trips to the other. The specific demand patterns of origin and destination along the day, yield a stable and mostly significantly increased trip distance, and came as a consequence of interaction between economically and socially active areas as those participating in Thessaloniki's case study.

Extending those observations, in Figure 187 is presented a more complete view. The information retrieval produces a great representation of trips. A deeper look in those clusters allows us to observe that there are many clusters that can cooperate. All those clusters in Figure 187 compose a complete area that is capable to perform ridesharing or micro mobility trips. All 6 clusters together are almost 50% of trips performed in Thessaloniki. Therefore, it seems that half of them performed around city centre and Kalamaria district. So, if a ridesharing service focus on that area it can capture the greatest amount of the demand. Another one observation is that those 6 clusters are paired in 3 mirror clusters. Thus, they can supply departure part as long as the arrivals as starting point. Additionally, near 40% of them has a trip distance up to 3km and can be performed by micro mobility. High demand also impose that it would be easier for a passenger or service managers to schedule a ridesharing trip. Current situation in that region dominated by car trips, and bus usage that usually leads to congestion events or overcrowded public transport routes. Therefore, those results promise that this mobility status can change and provide great opportunities on many transport services.

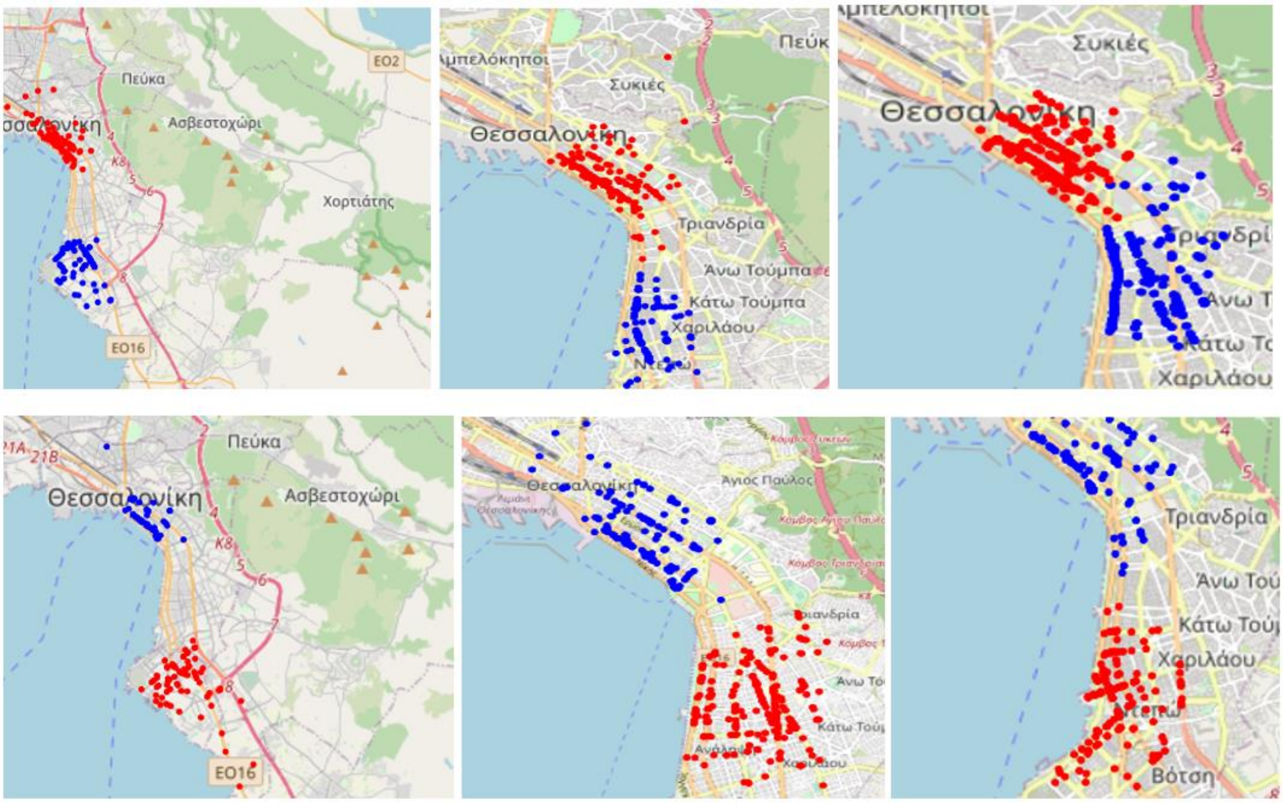


Figure 187. Super cluster. (Blue are the origin points, Red the destination)

3.4.4. Validation of the structural similarity measure

As with Section 3.1.6, the purpose of this part is to validate the structural similarity measure of OD matrices proposed in MOMENTUM and described in Section 2.3. in a new benchmark using data from the Thessaloniki case study. Concretely, we will use OD matrices extracted from FCD data from taxi trips, that is, the same data source use in the two previous sections. The next part of the section is structured in an analogous way to Section 3.1.6. We will start with the validation approach used, after that, we will comment on the results, and finally, we will also discuss the main conclusions drawn from this analysis.

3.4.4.1. Validation approach

For the validation in the case of the use of Thessaloniki, the approach used is very similar to the one presented in Section 3.1.6, although with some differences. Below we detail the day-to-day comparison approach used, the objectives set, and the OD matrices that have been employed in the validation:

- **Approaches for day-to-day comparison:** in this case, we have only used the pairwise-daily approach. As we will comment later and as already seen in Section 3.4.1, the OD matrices generated from FCD from a subset of the trips of half of the taxi fleet of the city, thus having a very low number of trips per hour and OD pair (in the vast majority of cases 0 or 1 trip) and are very dispersed. This generates a lot of noise when considering short time intervals such as one hour, which makes it difficult to compare them. For that reason, in this case, we have chosen only to compare aggregate daily OD matrices.
- **Objectives:** the objective here is to validate the structural similarity measure on another scenario and also on other types of OD matrices, with a low number of trips and therefore more dispersed (higher number of zeros in the OD matrices). To do this, we will use similar patterns to the ones set out in Section 3.1.6, with the only difference that we expect Mondays and Wednesdays to be similar due to the fact that shops are closed on the afternoon (it will be the new pattern (2)), while during Tuesdays and Thursdays they are open so these two days may also have important similarities (it will be the new pattern (3)). In addition, since these are taxi trips, their purpose is less related to a daily commuting trip between home and work and more to leisure and other daily activities. Therefore the results may differ significantly from the ones in Madrid.
- **Input OD Matrices:** to validate the structural similarity measure for OD matrices, we use a zoning that divides the Thessaloniki metropolitan area into 315 traffic zones. In this way, the compared matrices have a dimension of 315x315. Regarding the periods, we used daily aggregated OD matrices for the whole 2019 year. In this way, a total of 365 OD matrices have been used for this analysis.
- **Implementation:** the implementation used was the same as the one described in Section 3.1.6.

3.4.4.2. Results

3.4.4.2.1. Comparison among days of the week

In this section we are going to verify the patterns (1) - (6), to see to what extent the proposed matrix comparison method is able to detect them with a different type of OD matrices than those used in the previous validation (taking into account the two new patterns) since they have a significantly lower number of trips.

To this end, Figure 188 shows the distribution of the structural similarity measure for each day of the week against the rest based on the intra-month comparison, using the pairwise-daily approach. Just to remind, each panel corresponds to one day of the week and each box to the distribution of similarity both with respect to the same day of the week and to the rest. In order to interpret the results, it is convenient to know that in the proposed measure, the lower the value, the higher the similarity between the compared OD matrices.

In order to avoid biases in the comparison, the following days associated with bank holidays have been removed for the analysis done in this section:

- January 2019: 1,2, 5 and 6
- March 2019: 11 and 25
- April 2019: 25,26,27,28, 29 and 30
- May 2019: 1
- June 2019: 16 and 17
- August 2019: 15 and 16
- October 2019: 28
- December 2019: 25,26 and 31

Observing the Figure 188 we can see that there is a strong difference between the weekdays and the weekends, which allows us to validate the pattern (1). However, unlike what was observed in Section 3.1.6, the days of the weekends with each other does not have high similarity. This is probably due to a very low number of trips on those days, which makes it difficult to find similarities.

The pattern (2) is also fulfilled in this case, as there is a high similarity between Mondays and Wednesdays. On the other hand, as regards the pattern (3), we also observe a high similarity between Tuesdays and Thursdays, although no big differences are observed with respect to other weekdays.

As far as the pattern (4) is concerned, we cannot see it in this graph, as Fridays do not seem to be different from the rest of the days of the week, but quite the opposite, as their similarity is higher. This may be due to the type of trips made by taxis, which are not regular commuting trips but more related to leisure or other punctual activities. In this way, maybe it is possible to not have very big differences between the trips done by taxis on Fridays compared to the rest of the weekdays. In any case, further research is required to investigate the possible reasons for this results.

Finally, the pattern (5) cannot be seen in the graph either, since except for the case of Fridays, the other days of the week are no more similar to themselves than to the rest. This shows that with this data, the measure is able to detect differences, but it finds it more difficult to detect similarities.

In Figure 189, we also show the distribution of similarity between the days of the week, but in this case separated by months as we also did in Section 3.1.6, to see to what extent these patterns are observed throughout the months. Here we can see that all months show a rough similarity between days of the week like the one we saw in Figure 188. That is, a big difference between weekends and weekdays (pattern (1)) and a big similarity between Mondays-Wednesdays and Tuesdays-Thursdays (patterns (2) and (3)). As said above, patterns (4) and (5) are not fulfilled in any month. As for pattern (6), this would be partially fulfilled, since although weekdays maintain their similarity throughout the months, the weekend days are not similar to each other in any of the 12 months studied.

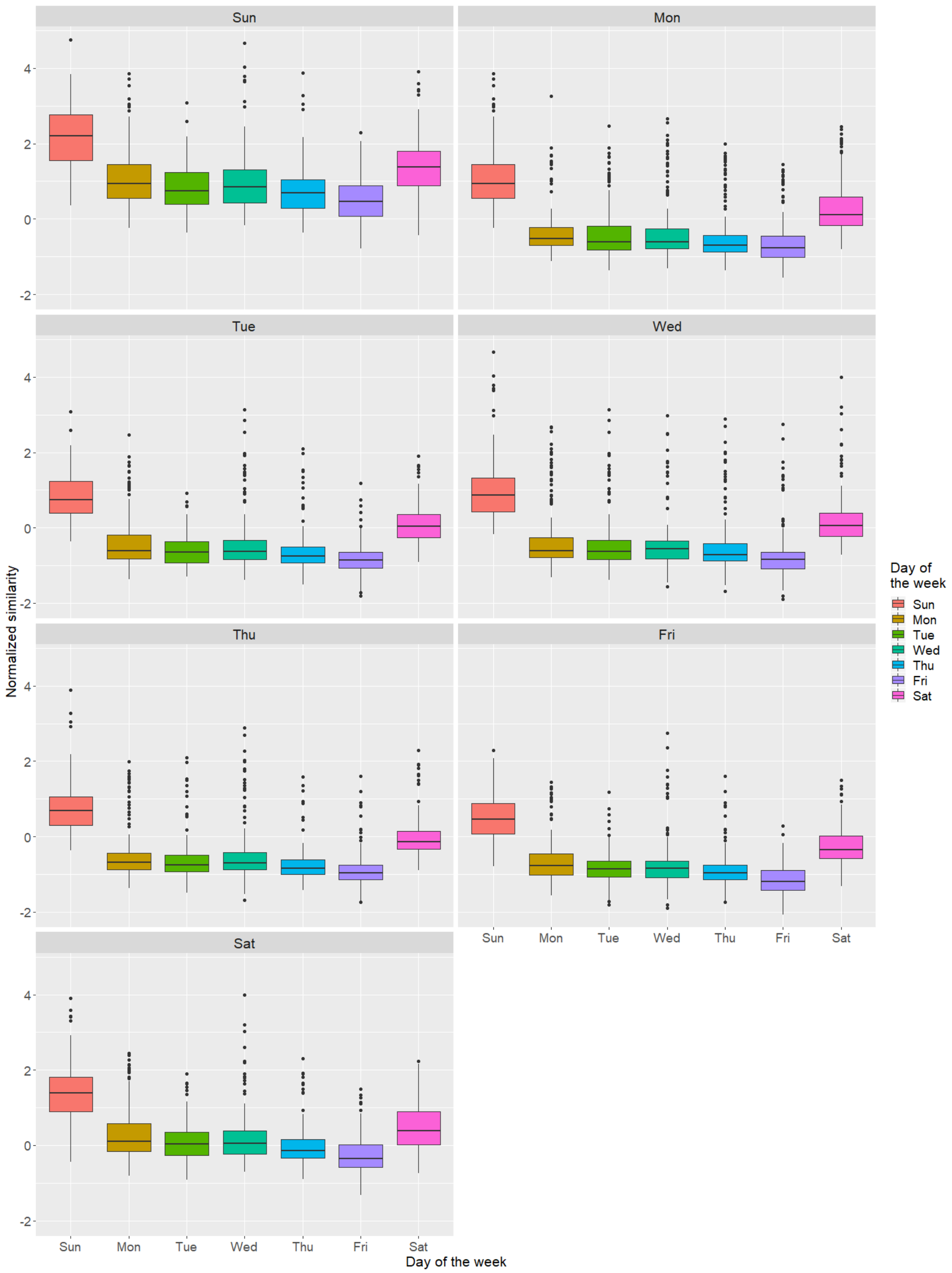


Figure 188. Distribution of the similarity for all days of the week (Approach pairwise-daily)

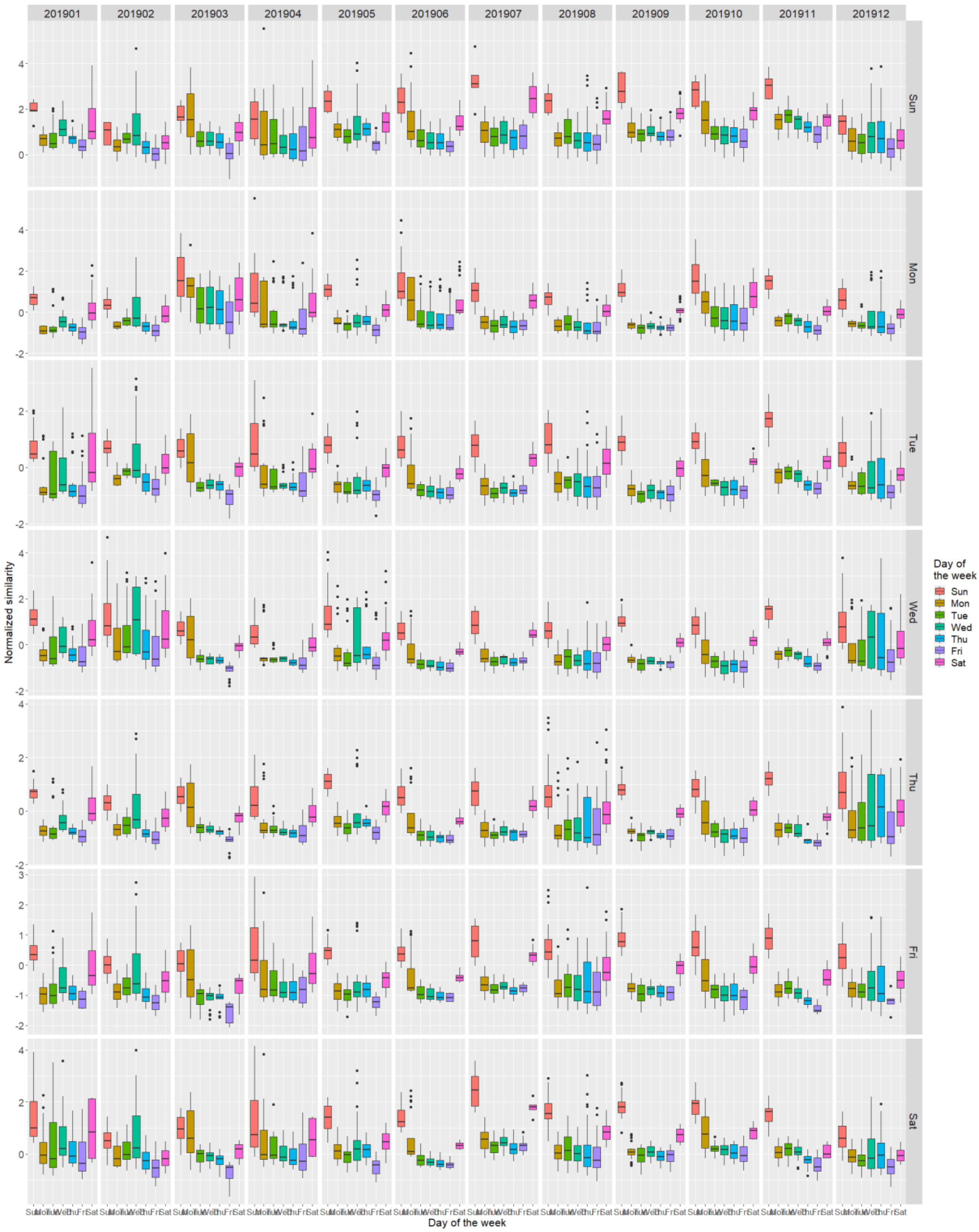
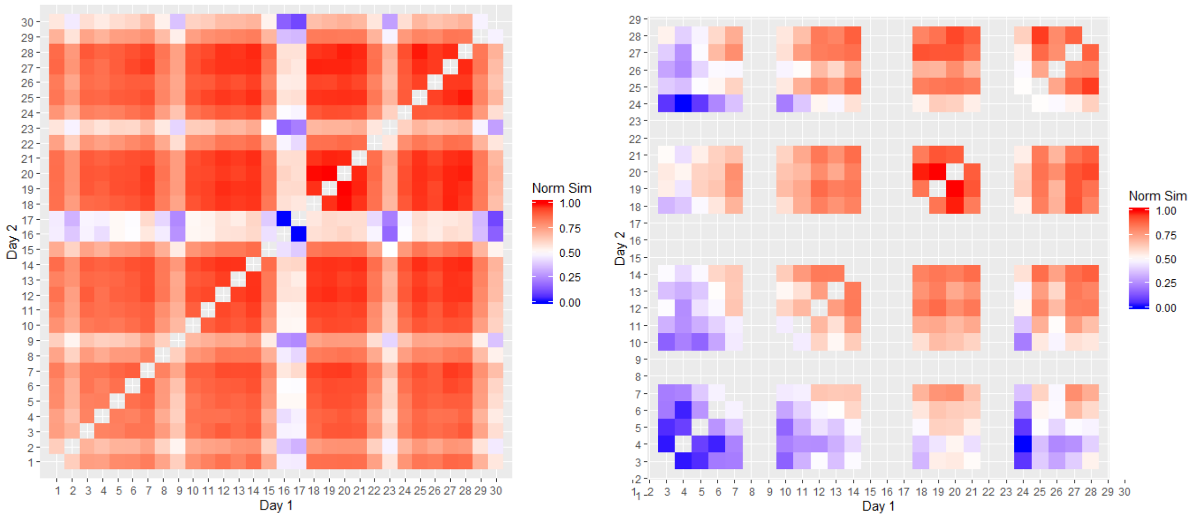


Figure 189 Monthly distribution of the similarity for all days of the week (Approach pairwise-daily)

3.4.4.2.2. Day to day comparison

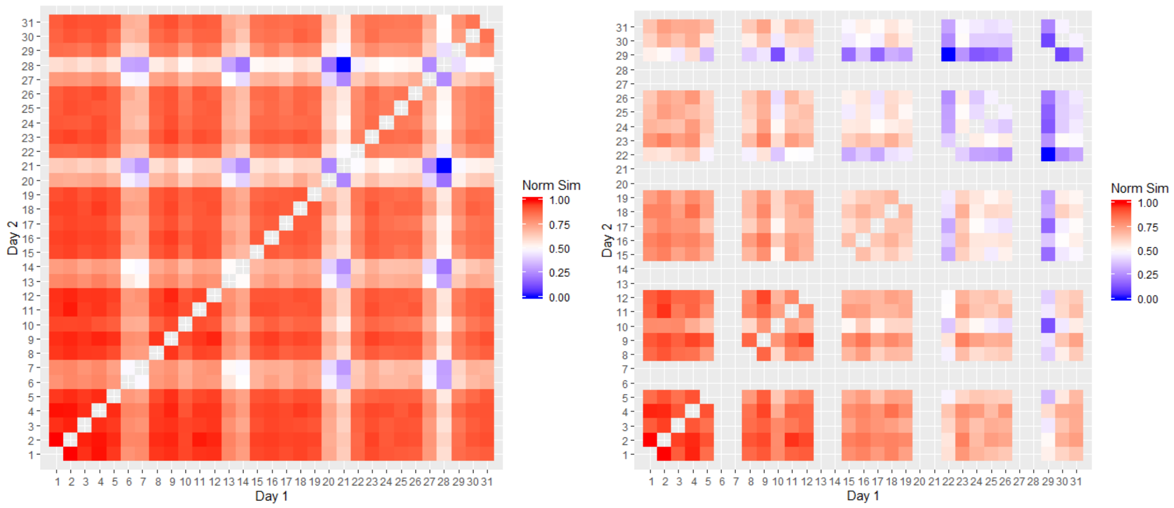
As in Section 3.1.6, we will use day-to-day comparisons to check whether patterns (7) and (8) are being met. Specifically, we will use the same type of figures as in Section 3.1.6.2.3 for the months of June, July, September and April 2019. To facilitate the interpretation of the figures, we will remind you that in these heat diagrams, the colour scale is given by the similarity between the days being compared. Dark blue colour indicates a lower similarity, white colour an intermediate similarity and dark red colour a high similarity. As the similarity values are standardised month by month, it is not appropriate to compare the colour scale from one month to another, nor between figures (a) and (b), as they may represent different similarity values.



(a) All days *(b) Only weekdays*

Figure 190. June 2019 with approach pairwise-daily

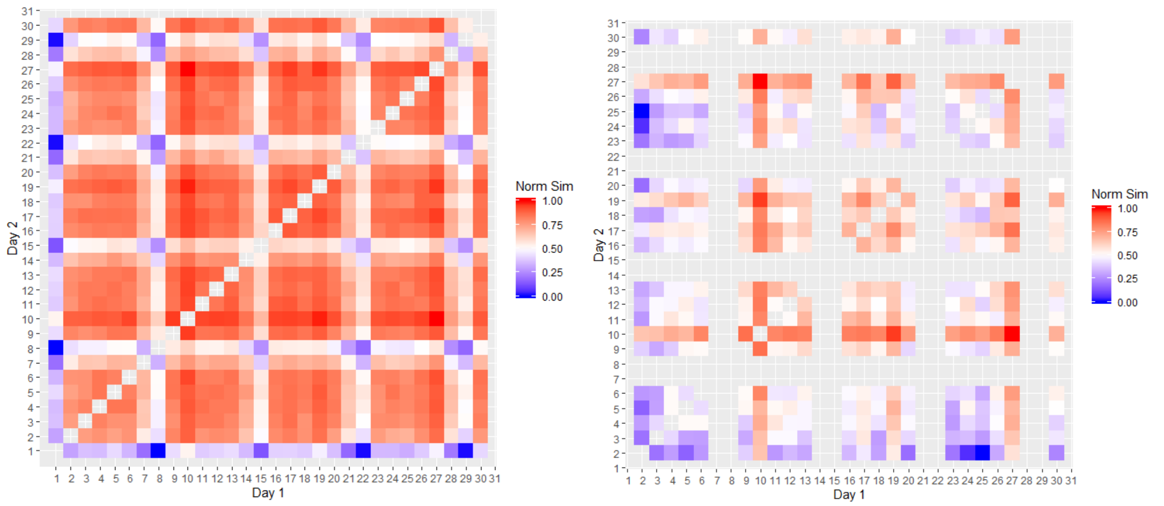
Figure 190, Figure 191 and Figure 192 allow us to validate pattern (7) as they correspond to transition month, concretely, to June, July and September 2019. If we look at figures (b) when only weekdays are displayed, we can see that there is a clear difference between the initial and the final part of the month. In June 2019 we can see that the first two weeks, and particularly the first one, are very different from the last two. In July this pattern is even clearer, although in this case, the dark blue colours that indicate less similarity appear mainly in the last half of the months. In the case of September, the differences are mainly observed in the first week, when usually a progressively higher number of commuters start of work after the return from holidays.



(a) All days

(b) Only weekdays

Figure 191. July 2019 with approach pairwise-daily

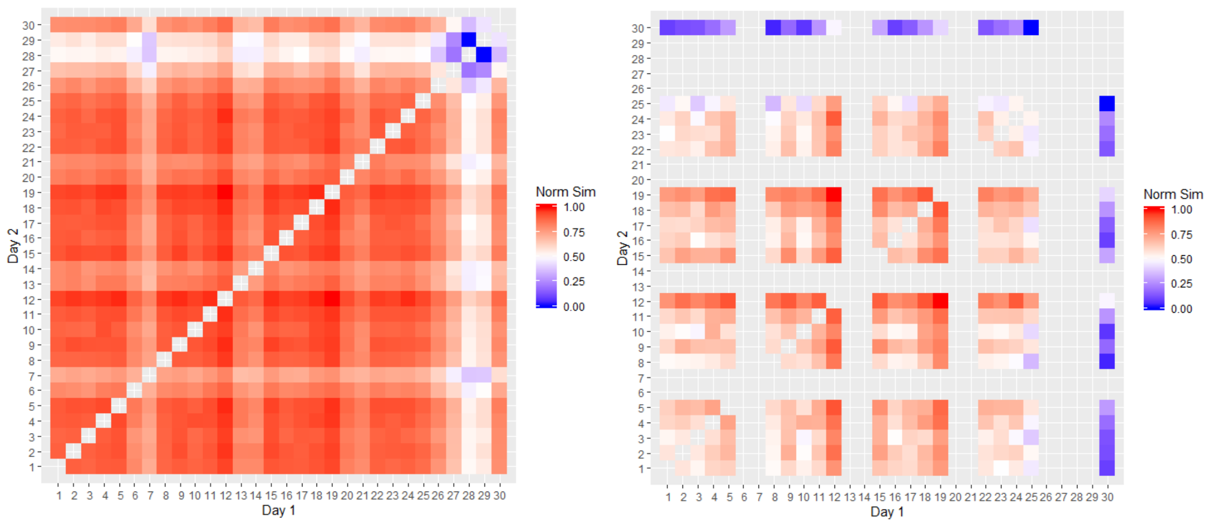


(a) All days

(b) Only weekdays

Figure 192. September 2019 with approach pairwise-daily

The pattern (8) is shown in Figure 193. This figure corresponds to the month of April 2019. At the end of this month, Orthodox Easter was celebrated with the 26th and 29th as public holidays. It should also be noted that May 1st is also a holiday. Therefore, the 25th and 30th of April can be considered as pre-holidays. If we look mainly at figure (b), we see that both the 25th and the 30th are quite different from the rest of the weekdays, especially the 30th. Furthermore, we see that the 25th is similar to Fridays, which is marked with a red colour, while the 30th is less dissimilar to Fridays than to the rest of the weekdays.



(a) All days *(b) Only weekdays*

Figure 193. April 2019 with approach pairwise-daily

3.4.4.3. Discussion and conclusions

As discussed in Section 3.1.6, there is a need for methods that allow for the comparison of OD matrices in an appropriate and efficient way in order to take better advantage of increased longitudinal data sources describing mobility patterns.

In this case study, the data source used to validate the structural similarity measure developed in MOMENTUM has been OD matrices extracted from taxi FCD in Thessaloniki. Unlike the OD matrices used in the Madrid case study, the ones used in this case register a much lower number of trips since they consider only a subset of the displacements made by half of the taxis in the city. In addition, the size of the matrix is larger, as a total of 315 traffic areas are considered, compared to 208 in Madrid. This also leads to a higher level of noise when similar mobility patterns are compared. This has allowed us to validate the similarity measure in another very different scenario. In this case, not all the well-known mobility patterns established to validate the measure were observed. It was found that the measure detected differences between weekdays and weekends and holidays, the similarity between Monday-Wednesday and Tuesday-Thursday), the similarity between the pre-bank holidays and the Fridays, and differences across weeks in transition months such as June, July and September. However, in this case, the measure was not able to capture the similarity of the weekend days with each other or the difference of Fridays with respect to the other days of the week. In our opinion, this last fact is due to the high noise these OD matrices with high granularity, and a low number of trips can present.

As we saw in Section 3.1.6 and in this one, the results of the structural similarity measure designed in MOMENTUM are promising and have a large number of applications. However, in order to improve its performance and robustness in a wide range of scenarios, the results obtained here also suggest future lines of improvement that we hope to undertake in the next stages of the project:

- The structural similarity measure developed is based on one-to-one comparisons of origins with origins and destinations with destinations. This can make it sensitive to errors in the zone allocation of the trips (e.g., edge effect) or to small fluctuations, particularly when the OD matrix has a very high granularity and a low number of trips. The main reason is that in these scenarios, the errors or fluctuations mentioned can represent a high portion of the total number of trips generated/attracted by the origin or destination compared. In this sense, a future improvement that we propose for the measure is to incorporate a group-to-group comparison approach of origins with origins or destinations with destinations, which will make it more robust in this type of scenarios.

- The measure does not incorporate geographical information on the location of origins or destinations. Including geographical information that allows the method to determine which regions are bordering or which destinations or origins are closer or further away from each other can help to improve its robustness. For this reason, it is also one of the future improvements we hope to incorporate.

4. Relation and implications of results for MOMENTUM Test cases

4.1. Implications for Madrid Case Study

The overall objective of Madrid case study is to include shared mobility impacts in the transport planning and decision-making processes. This is explored through the development of new analytical solutions that will complement and enhance the existing multimodal macroscopic transport model developed by EMT Madrid. The focus of the case study is on the effects that shared mobility has in the travel demand patterns across the city and on the implications for EMT Madrid as a major public transport operator in the city. The data analyses conducted in this WP3 have unveiled some trends in relation to shared mobility use and adoption that are relevant to the research questions posed for the case study in the deliverable D2.2.

The analysis of the mobility household survey and the customer data from two shared mobility services allowed us to identify which profiles are more prone to adopt new mobility solutions in the city. The results do not deviate from what is described in the literature from other cities: bias towards **male young adults**, with **higher education and income levels**. Some patterns differ from one mode to another: **some affordability barriers were observed for the case of moto sharing that were not observed for bike sharing**. This raises a viability concern for the deployment of shared mobility services in lower income areas beyond the central districts that deserves a closer monitorisation. The use of the analysed shared mobility services is characterised by rather **short-medium distances** (average trip distance below 3 km), but it has to be acknowledged that car sharing services are likely to have larger averages. **Bike sharing services are used to some extent in combination with public transport**, as the results from the household survey show than 22% of the uses were part of multimodal trip chains. However, this is not the case of car sharing and moto sharing. **Most users (67%) declared to have a private vehicle available when they performed a shared mobility trip**. This percentage exceeds by 25 points the one for public transport. This suggests that shared mobility services may be attractive enough for some citizens to prevent them from using their private vehicles.

The implementation of an UVAR zone restricting private car access to the city centre offered a great opportunity for assessing the potential of shared mobility services as a substituting mode to the use of private vehicles. The combined use of several longitudinal data sources (traffic counts, mobile network data and shared mobility operation data) and the mobility household survey has allowed us to quantify the variation of traffic flows due to the implementation of the UVAR zone, to check if this policy measure was followed by modal shifts and if shared mobility services were able to capture part of the demand that was once travelling by private vehicles. The implementation of the UVAR zone and the related measures to reduce road capacity and increase pedestrian and cycling space **reduced traffic levels by 13% in the entries and exits to the area and by 4% in its perimeter** (2019 vs. 2017). According to the data, **less people travelled to the city centre** in the months immediately after the UVAR zone implementation, but it is likely that some modal shifts happened given that there **no relation was observed between the decreases in the volume of trips to the city centre from each district or at each hour with the modal share of private vehicles** registered in the household survey for those flows. When looking at the data from a bike sharing service and from a moto sharing service, it is found that it is rather unlikely that the bike sharing service would have attracted trips previously made by car. However, the yearly comparison between October 2018 and October 2019 suggests that the **moto sharing service may have captured trips to/from the city centre from private vehicles**. This comes in addition to the findings from the mobility survey related to the availability of a private vehicle when performing a shared mobility trip.

Finally, the availability of a recent mobility household survey has allowed us to **validate and contrast the results of prominent big data sources for describing mobility patterns in the city of Madrid** (mobile network data and operation data from shared mobility services). The rich information provided by travel surveys is often too expensive and resource-hungry to maintain it as updated as required for monitoring rapidly evolving mobility patterns. It has been proven than the information provided by mobile network data help overcome these limitations, by providing a very similar image of trip generation and distribution in a metropolitan environment such as Madrid region. In addition, this source enables to perform comparison of seasonal and intra-week patterns. EMT Madrid and other authorities with responsibilities in mobility planning and management are already incorporating this data source as part of their tools for monitoring travel demand. These agents **will take advantage of the methodologies developed by MOMENTUM to process these data (similarity measures and representative OD matrix estimation)** that have been demonstrated in the context of the case study. The case for operation data from shared mobility services is similar. It has been observed that household mobility surveys may be useful for characterising some adoption patterns of these services, given that the required sample for extracting valuable insights about user profiles is rather small. However, the characterisation of use patterns of emerging mobility services from these surveys is challenging, due to the limited number of trips that can be captured through these surveys. The exploitation of the **operation data has allowed us to characterise the most relevant variables about the use of shared mobility services** in Madrid city.

4.2. Implications for Leuven Case Study

As indicated in deliverable D2.2, the main objective of the City of Leuven in the Test Case proposed for the MOMENTUM project is the further regional integration of all different transport systems and mobility providers with the future spatial developments. With this aim, the case study establishes three policies to be analysed:

1. The implementation of a new circulation plan for districts within the city such as Kessel-Lo and Heverlee.
2. The introduction of 50 so-called ‘Mobipunten’, i.e., small to medium-scale mobility hubs with car-sharing, bike-sharing or scooter-sharing facilities, parking infrastructure and charging stations.
3. The implementation of Regionet, a new overarching strategic mobility plan[9]. In particular, the introduction of bus lanes and the implementation of high-quality bike infrastructure, in combination with pricing policies and the implementation of peripheral parking and autonomous shuttle buses will be tested.

The analysis of policies 1 and 3, due to their characteristics and the availability of data in this use case (see Deliverables D3.1), is linked to the new multi-modal traffic model that will be developed in work packages WP4 and WP5. For this reason, in this deliverable, we will only analyse the implications of the results shown in Sections 3.2.2 and 3.2.3 for policy 2, the introduction of the 50 "Mobipunten". The specific questions raised for this policy were the following:

1. What is the optimal size of the shared fleet at these ‘Mobipunten’?
2. How will they affect car ownership and modal split?
3. How many transfers from one mode to another can we expect at these points?
4. How could subsidies for mode combinations impact this?
5. In a system where shared bikes can be returned to any station; how much intervention would be needed to align asymmetric supply and demand at different locations and at different times (because of large events for example)?
6. Would the related cost be offset by the extra efficiency and use, compared to a system where the shared bike has to be returned to the original station?

[9] See: http://regionetleuven.be/

As with policies 1 and 3, the final answer to these questions has to be made through the multi-modal simulator mentioned above, but the results analysed in Section 3.2, do have implications in a more direct way for question B, on how Mobipunten will affect car ownership and modal split.

We will first focus on car-ownership and the implications that car-sharing can have on it. As we have seen in Sections 3.2.2 and 3.2.3, those people who have a subscription to car-sharing services or who declare that they are already doing car-sharing have on average fewer cars and more bicycles. These results are in line with the literature on the subject. For example, Martin and Shaheen (2011) and Martin et al. (2010), show a decrease of almost 50% in vehicles owned by the households of 6281 respondents who were car-sharing members. Katzev (2003) also found a causal relationship between car-sharing membership and a decline in vehicle ownership. Specifically, they reported a 44%, 60% and 25% reduction in car-ownership in the Netherlands, Switzerland, and Canada, respectively. More recently, in Le Vine and Polak, 2019, the authors found that 37% of free float car-sharing users in London claimed that their ownership of private cars had had an impact, and furthermore, of this 37%, 83% indicated that because of this service, they had stopped buying a car that they would otherwise have bought. In the same vein, but in the city of Basel (Switzerland), Becker et al., 2018 claim that 6% of car-sharing customers reduced their ownership of private cars due to their membership. Finally, another recent study carried out in 11 European cities using data from the company SHARE NOW, found a similar effect, although with considerable differences from one city to another. They found, for example, a ratio of cars sold per free-floating car-share (FFCSV) ranging from 2.1 in Madrid or 5.3 in Lisbon, to 7.8 in Vienna or 18.6 in Copenhagen. Although all these studies show very positive results, other authors, such as Zhou et al. 2020, argue that they may be very optimistic, as they may have selection biases since they are mostly based on the preferences of car-sharing members. In fact, through a survey of the general public in Australia, they find that there is no significant impact of car-sharing membership on vehicle ownership.

As we saw in Section 3.2.4, car sharing seems to be related to the number of private vehicles in those higher-income households where members are doing car sharing or have subscriptions to this type of services. The fact that the two variables related to car-sharing appear as one of the main factors predicting car-ownership among higher-income households offers indications that it may be affecting car-ownership given that are households who can afford it. However, as already mentioned this hypothesis is weak because of two reasons: on the one hand, with the data available we cannot assess the direction of the causal relation (car-sharing implies less car-ownership or vice versa), and on the other hand, we cannot rule out a selection bias such as that noted in Zhou et al. 2020.

In any case, if there is some impact in car-ownership, we have also seen that is still small, which may be partly due to the still low penetration of this type of service, as only 12% of the population claimed to be doing car-sharing in 2017, and only 4.5% of households had a subscription to these services. If Mobipunten increases and makes car-sharing services more accessible and therefore increase the level of penetration of these services, in principle, maybe the impact of car-sharing on the number of private vehicles in this city will be more significant than observed so far. However, the results shown in Sections 3.2.2.2.13 and 3.2.3.2.13 indicate that this impact could take time or may not occur, as it was observed that the availability of cars in households in a given area is inversely correlated with the use of car-sharing in that area. Thus, the data show that the deployment and design of Mobiputen should take into account the profile of the people and households of each area. In any case, it will be in WP5 where a more definitive answer to this question will be given.

As for the second part of the question, relating to modal split, the influence of car-sharing on this is also a topic that has been addressed in the literature. In this sense, there are several works that report benefits in terms of increased use of soft modes due to the irruption of this car-sharing service, although these effects vary from place to place and also depends on the type of service (Jackobson et al. 2018, Martin et al. 2016, Ceccato 2020). In the specific case of Leuven, as we have seen above in Sections 3.2.2 and 3.2.3, people living in households with at least one car-sharing subscription or who are already doing car-sharing, have, on the one hand, a lower number of private vehicles in the household, and on the other hand, a higher probability of using soft modes to travel to work. Although this may lead to think that a higher offer of car-sharing services may lead to greater use of soft modes, as we mentioned in Sections 3.2.2.2.8 and 3.2.3.2.8, indeed car-sharing might be competing with public transport. For this reason, wider accessibility to car-sharing services could have a negative impact on the use of public transport. However, the data suggest that this competition is low. On the other hand, the fact that households without a car and younger people make more use of car-sharing services may also suggest that these services are discouraging private car purchase, and this may contribute in the future to increased use of soft-modes. However, as mentioned above, the still low penetration of car-sharing services and a possible selection bias in the results are important limitations to reach firm conclusions, so we suggest taking the results with caution. Furthermore, as with the previous question, a definitive answer will be given in WP5, since this requires the use of transport simulation systems.

4.3. Implications for Regensburg Case Study

As described in the deliverable D2.2, the main objective proposed for the Regensburg Test Case in MOMENTUM is to have a reliable model for deriving scenario outputs and guiding evidence-driven policy input by integrating the new mobility services into the existing model. According to this main objective, the questions raised were the following:

1. Vehicle automation opens the room to innovative transport supply schemes. To what extent can the autonomous people mover complement conventional public transport services?
2. Shared mobility services claim to reduce the need for owning a car. To what extent the shared mobility services implemented in the city have an impact on car ownership?
3. Air pollution is at the top of policy priorities affecting urban mobility. What is the impact of the new mobility options on air quality?

The answers to questions A and C, due to the type of analysis they require and the data available, need to be addressed through the traffic simulation model that is intended to be built for this city in the context of Work Packages 4 and 5.

As mentioned in the previous section, the impact on car-ownership of shared mobility services and in particular car-sharing has been studied in the literature, and most of them find a positive impact on it, leading to a lower number of private vehicles at home. However, it has also been observed that this impact can be variable from one country to another or from one city to another. In the specific case of Regensburg, we did observe that those individuals using car-sharing services tended to have significantly fewer cars in their homes, regardless of how often they used them, although this reduction was greater in those cases when individuals interviewed used these services with a high frequency (1-3 days per month or 1-2 days per week), which somewhat suggests that their impact has been positive.

However, as we saw in the last section, there are important limitations in this study: first, the available data that does not allow to establish the specific direction of the causal relation (e.g. car-sharing implies less car-ownership or incentivize the use of a car by people who do not plan to buy a car); second, there is a possible selection bias in the data used in this study; and third, the low penetration level of this service make difficult to generalize any conclusion. For this reason, it is important to take these results with caution. In any case, the final answer to this question will also be addressed within WP5.

Other implications of the results relevant for WP4 and WP5 is the fact there is no data showing daily use or even frequent use per week (e.g. 3 to 4 days per week) since as we saw before the most regular car-sharing usage group in Regensburg was 1-2 days per week but with an almost residual representation. Hence, it may not be possible to adequately model this system in the traditional transport models, as the kind of usage would not be usually accounted for in the origin-destination matrix. In fact, in the household survey used in the analyses, only 2 out of 8631 trips used car-sharing as a transport mode.

4.4. Implications for Thessaloniki Case Study

The overall objective of Thessaloniki case study is to improve the planning and decision-making process for the introduction of resilient sustainable mobility schemes, with main emphasis in adoption of DRT, ridesharing and vehicle sharing (micromobility, bike and electric car) mobility solutions towards MaaS in the agglomeration. This will serve also to develop techniques which can facilitate proofs of concept of new mobility schemes and the to improve and extend the use of innovative data sources (Floating Car Data, point-to-point detections, social media, etc.) in the transport modelling process.

Regarding the implementation of the DRT and their contribution to the sustainable mobility, they are expected to play a role in the surrounding of the agglomeration, where the population density is low. This assumption is validated through the case study of the municipality of Thermi described in Subsection 3.4. Indeed, Thermi which is characterized as a low-density area with low level of public transport services, it seems more suitable or the implementation of a DRT line through the KPIs calculation. As concluded in subsection 3.4.3, *a **12 stops DRT line with two vehicles of 20 passengers capacity each** will **serve over 62%** of passengers **in Thermi** earlier and more reliable than bus line. To get a **greater share of 80% or more** it needs **more than 4-5 vehicles**.*

Once a district is characterized as suitable for DRT services, the DRT planning methodology described in Subsection 2.2 is used for the planning and definition of the operational schedule of a DRT line. Through the methodology the main characteristics of the DRT line such as the number of vehicles needed, the frequency and the capacity of the vehicles are defined. The second mobility scheme to be tested in Thessaloniki as it is mentioned in D2.2 is the ridesharing service and its role in the transport system of the city. A ridesharing has been already implemented in the framework of the GALILEO4MOBILITY project in 2019, providing taxi-sharing services to people living in 2 areas (one suburban and one peri-urban) towards the city center and back home. KPIs calculation in Subsection 2.2 shows that the municipality of Kalamaria is more suitable for ride sharing services. Indeed, the service developed through the GALILEO4MOBILITY project worked well in Kalamaria while in Thermi it was not successful. As concluded in Subsection 3.4.3, **the 6 clusters created through** the trajectory similarity measure **are almost 50% of trips performed in Thessaloniki**. Therefore, since it seems that half of them performed around city centre and Kalamaria district, if a ridesharing service focus on this area it can capture the greatest amount of the demand.

The new methodology for the demand clustering for ride-sharing, as it is described in the Section 2, could provide answers to the main questions related to the design of the ride sharing service mentioned in the test case in Deliverable D2.2, contributing to the extension of the ride sharing service to the city areas where its success is more feasible. The clustering of points in the same direction through the use of the matching algorithm during the route analysis process will provide useful insights for the areas where a significant number of origins and destinations points are concentrated as well as for the number of people per vehicle. Thus, these areas could be characterized as "areas of high potentiality" for ride sharing trips. The distribution of routes during the day and the week will also contribute to the scheduling of the ride sharing lines. The identification of the "areas of high potentiality" for ride sharing trips and the hourly distribution of the lines during the day could be a valuable input for the creation of a ride sharing lines network that not only offers the possibility of door2door trips between the suburban areas and the city centre but also feeds public transport in areas with low level of connectivity.

The third test case of bike-sharing and micro-mobility and their impacts in transport planning will be investigated through the modelling processes of WP4. Some first results are extracted in Subsection 3.4.3 in which we observed that **near 40%** of the trips in the **6 clusters** has a **trip distance up to 3km** and thus they can be performed by micro mobility modes. The modelling will focus on the relation between micro-mobility and cycling infrastructure, by exploring the potential contributory factors for the bike lanes network planning. In addition, characteristics such as the fleet size, bike-sharing model (dock or dockless), distribution and rebalancing operations of the current and planned bike-sharing schemes will be evaluated. Vehicle-sharing schemes are expected to perform better between the city centre and the suburban and peri-urban areas for electric cars, while for bike sharing and micro-mobility it may work better if limited to the city centre and the suburban areas only.

Thus, the new methodologies developed combining maps analysis, KPIs and route analysis will contribute to the development of a decision-making tool in WP5 towards the characterization of an area as more appropriate for the implementation of a ride sharing or a DRT service.

5. Conclusions

In this deliverable, we have presented the methodologies and data analyses carried out in the WP3 of the MOMENTUM project to fulfil the final objective of this work package: to develop new data fusion and artificial intelligence algorithms for the extraction of the mobility patterns and travel choice behaviour drivers of different population groups. The design and development of these new methodologies and data analyses have been designed taking into account the test cases described and collected in Deliverable D2.2 “Specification of the MOMENTUM Test Cases” and also the requirements from WP4 and WP5. These have been validated using the datasets collected in the first stage of this WP3 and documented in Deliverable D3.1. Having said this, **the main outcomes of this deliverable are** the following:

- **New methodologies based on data fusion and Artificial Intelligence for extracting and analysing mobility patterns**. More specifically, the methods developed have been the following:
    - **A new method for estimating a representative OD matrix** for a certain period of time by the fusion of a set of similar OD matrices using a variant of the primal Support Vector Machine problem. The method has been validated using OD matrices extracted using FCD data from a taxi fleet in Thessaloniki and OD matrices extracted from mobile phone network data in Madrid. The results showed that the obtained representative OD matrices have greater similarity to the real data than those obtained through the classical approach of calculating the mean OD matrix. However, it was also observed that this method led to a significant increase in the number of trips compared to the average matrix, so it has to be further explored.
    - **A novel structural similarity measure for OD matrix comparison** based on a new concept of OD matrix structure whose comparison can be done very efficiently using new High-Performance Computing paradigms such as computational graphs. The proposal has been validated by analysing different patterns that are well known to traffic practitioners, and that should reproduce the structural similarity measure (e.g. different travel patterns between weekdays and weekends). For this purpose, OD matrices with different characteristics from two case studies have been used. On the one hand, in the Madrid case study, where we used OD matrices extracted from mobile phone network data and, on the other hand, on the Thessaloniki case study where we employed OD matrices extracted from taxi FCD data. Furthermore, two approaches were considered for the validation of the above-mentioned patterns, based on the comparison of daily and hourly matrices, respectively. The experimentation carried out in the Madrid case study confirmed that the proposed measure of structural similarity was capable of capturing practically all the patterns mentioned above, thus ensuring its validity. Regarding the Thessaloniki case study, most of the patterns were captured by the measure, but the noise present in the OD matrices (much sparser than in the Madrid case study) avoid to capture all of them. Furthermore, we have seen that this new measure, due to its fast computation, makes possible the quick characterisation of mobility patterns through long periods. This enables, on the one hand, to select the most appropriate set of dates and times to obtain representative matrices, thus complementing the previous approach, and on the other hand, to select representative days to carry out fieldwork such as surveys or counts. Another possible application would be the detection of outliers in the generation of OD matrices. Apart from this, different lines of improvements were discussed in order to make the similarity measure more robust when dealing with very sparse OD matrices.
    - **A new framework for estimating the potential demand that can be served by DRT or Ride-sharing systems.** The new framework has three different stages: 1) the data with disaggregated demand data is pre-processed to extract the probability distribution of each DRT stop; 2) the stops to be served by the DRT line are defined using k-means and savings algorithm, and; 3) the operation required to serve the stops is estimated using the Integrated Dial a Ride problem. This

deliverable encompasses the two first stages, whereas the third one will be addressed in WP4. The two first stages have also been validated using FCD data from taxis in two regions of Thessaloniki, Thermi and Kalamaria. The results show that Thermi, which is characterised as a low-density area with limited accessibility by public transport services, seems more suitable for the implementation of a DRT line, whereas the municipality of Kalamaria, more densely populated and better connected with public transport, is more suitable for Ride-sharing services.

- **A generic approach for extracting user trip diaries from shared-mobility operation data.** The method aims at standardising the processing of operation data from shared-mobility operators by the creation of an entity (user or vehicle) diary. The diary contains the all the trips performed by an entity during a fixed period of time together with the date and the user identifier, and additional information if needed (e.g. age, gender, weather, etc.). The approach has been successfully validated using data from a dock-based bike-sharing system (BiciMAD) and a moto-sharing service (Muving) in the Madrid case study.
- **New methods for home location and income level inference from shared-mobility operation data.** Regarding home location inference, three different approaches have been followed for the design of the new methodology, based on cross-sectional data, longitudinal data and Machine Learning. The method has also been validated using BiciMAD and Muving operation data. The results obtained showed that the approach that leads to a better performance was the one based on longitudinal data because it provided richer information. However, the low accuracy of the methods based on recurrent use patterns revealed that many users do not take shared modes on a regular basis. As for the income level inference method, two approaches have been followed for its design, one based on customer databases with home location available, and another one based on trip data with home location information available. The latter methodology was validated using the user database from Muving, which contains the home location of each user. Results showed that both approaches to income distribution inference, trip-based and user-based, are consistent and yield similar results, hence being possible to infer the income distribution of a service from a collection of trips from various days.

- **New studies concerning emerging mobility solutions and the use of prominent big data sources in the MOMENTUM case studies** These studies focus on the adoption and use of shared mobility services, and the use of big data sources to describe mobility patterns and to assess the effects of certain policies. The following is a list of these studies and their main conclusions:
  - **Comparison of household survey data and mobile network data.** For the Madrid Case Study, two OD matrix estimation approaches were compared, one based on a mobility household survey and one based on mobile network data. The comparison of both methods has been performed over different features of the matrices (e.g. volumes of trips, hourly distributions of distances and additional segmentation based on sociodemographic and trip characteristics). The results show that both OD matrices have demonstrated high degrees of similarity in most dimensions (although there were some discrepancies in some of them), suggesting that both alternatives provide reasonable estimations of travel patterns of the regional OD matrix.
  - **Shared-mobility role in restrictions to private vehicles.** Again, for the Madrid Use Case, a series of experiments were carried out to study the impact of emerging mobility solutions in Madrid and also about the technical opportunities and challenges that the use of big data sources entails for such assessments. These studies are based on the impact of the implementation of the UVAR zone 'Madrid Central' that restricted private car access to the city centre. Several longitudinal data sources (traffic counts, mobile network data and shared mobility operation data) and the mobility household survey were used to check if this policy measure was followed by modal shifts and if shared-mobility services were able to capture part of the demand that was once travelling by private vehicles. The study showed that the implementation of the UVAR zone and the related measures to reduce road capacity and increase pedestrian and cycling space reduced traffic levels by 13% in the entries and exits to the area and by 4% in its perimeter (2019 vs 2017).

Furthermore, the analysis revealed that fewer people travelled to the city centre after the UVAR zone implementation, although probably some modal shifts happened. The study of the data from bike-sharing and moto-sharing showed that it is rather unlikely that the bike-sharing service would have attracted trips previously made by car. However, the moto sharing service may have captured trips to/from the city centre from private vehicles.

- **Shared-mobility adoption.** Within the WP3 of MOMENTUM, the socio-demographic characteristics of users adopting shared mobility services have been studied in three of the four case studies, specifically in Madrid, Leuven and Regensburg. The main conclusions for each case study are as follows:
    - **Madrid Case Study.** This study was carried out, on the one hand, using microdata from a household survey, and on the other hand, using operational data from BiciMAD and Muving. These analyses showed that the profile of the users of this type of service is biased towards young male adults, with higher education and income levels. Some patterns differ from one mode to another: some affordability barriers were observed for the case of moto sharing that were not observed for bike sharing. Most shared-mobility users (67%) declared to have a private vehicle available when they performed a shared mobility trip. This percentage exceeds by 25 points the one for public transport. This suggests that shared mobility services may be attractive enough for some citizens to prevent them from using their private vehicles.
    - **Leuven Case Study.** In this case, the adoption of shared mobility services was studied using data from a liveability survey in Leuven. The analysis considered two perspectives: on the one hand, that of people living in households with at least one subscription to car-sharing services and, on the other hand, the car-sharing willingness of the population. Regarding the first of the two analyses, the results showed that people living in households with at least one car-sharing subscription are mostly young, highly educated and employed, but unlike Madrid and Regensburg (and most studies in the literature), with a higher bias towards women. They also tended to make higher use of so-called soft-modes (e.g. bicycle and public transport) to get to work. Apart from this, the households in which these people lived had a bias towards high-income levels, more members and fewer private cars at home. Regarding the second study, which analysed the car-sharing willingness of individuals, the conclusions drawn were similar to the first one, although with some interesting differences. The people already doing car-sharing were mostly young and employed women, although with a relevant percentage of students (almost 20%), and with a high level of education. Similar to the previous analysis, these individuals were more likely to use soft modes to go to work. In terms of the households in which these people lived, an interesting observed difference is that the distribution grew at the extremes (higher proportion of people doing car-sharing in the lower and higher ranges of household incomes and lower in the intermediate ones). This shows that in Leuven, car-sharing may be reaching people with less purchasing power, unlike what was observed in Madrid with the moto-sharing service. As for the remainder characteristics, the average household size was smaller, they had fewer private cars and a higher number of bicycles when compared to households of people not willing to use car-sharing. Another interesting aspect of this analysis was the relation between car-sharing adoption and car-sharing supply. We observed that the adoption is positively correlated with the number of car-sharing vehicles available in each district. Nevertheless, it had low explanatory power when predicting the number of people living in households with at least one car-sharing subscription or the people that are already doing car-sharing. However, when the car-sharing supply was combined with the number of households with no car in each district, these two variables explained 74%

and 86% of the variance of people that are doing and that want to do car-sharing, respectively.

- **Regensburg Case Study.** The study of the adoption of car-sharing services in Regensburg was also done using microdata from a mobility survey. In this case, there were also quite a few similarities with the two previous studies, but also relevant differences. The individual profile of people who reported using car-sharing services, as in Madrid, was more biased towards men. In terms of age, work status and education, they followed the same patterns as in the two previous case studies: young, employed and highly educated. In terms of mobility, as in Leuven, these people tended to make more use of the bicycle and public transport, and less use of the private car. Finally, in terms of the characteristics of the households in which people who use car-sharing live (similar to the second study in Leuven), it was observed that low and very high income households are over-represented, and those with medium incomes are under-represented, when they were compared with the households of people that are not doing car-sharing. Concerning their size, unlike Leuven, in Regensburg, they tended to be mostly single-person households. Finally, in terms of the private car and bicycle ownership, the bias was similar to that of Leuven, with a greater tendency towards lower and higher ownership of cars and bikes, respectively.

- **Shared-mobility use.** Similar to the adoption of shared mobility, the use of these emerging mobility modes was also studied for the two use cases where there were operational data of such services, namely Madrid and Regensburg. The main findings are detailed below for each case study:
  - **Madrid Case Study.** In this case study, the use of the two mobility services for which there was operational data, BiciMAD and Muving, was analysed. The results showed that these services were used for short and medium trips, with an average distance of less than 3km. Furthermore, it was observed that BiciMAD, the public bike-sharing service in Madrid, was used to some extent in combination with public transport, as the results from the household survey showed that 22% of the uses were part of multimodal trip chains.
  - **Regensburg Case Study.** In this city, operational data from two car-sharing services were analysed, one deployed in the urban environment of Regensburg and the other in the district of Regensburg. The results showed that these services were mostly employed for medium and long trips, with an average travel distance of 25 and 45 km for both services, respectively. The length of the trips, the patterns of the daily and hourly start of the trips, as well as the time in which they were booked in advance, suggested that their use was rather spontaneous and sporadic, and that were probably mostly linked to personal business, leisure and shopping activities. However, it was noted that there was a small portion of trips that could be linked to work activities.