Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 37 (2019) 417–424

Transportation Research Procedia

www.elsevier.com/locate/procedia

21st EURO Working Group on Transportation Meeting, EWGT 2018, 17th – 19th September 2018, Braunschweig, Germany

Fusing mobile phone data with other data sources to generate input OD matrices for transport models

L. Montero[a*], X.Ros-Roca[b], R. Herranz[c], J.Barceló[b] †

[a]*Universitat Politècnica de Catalunya, Jordi Girona 1-3, Barcelona 08034, Spain*
[b]*PTV IBERIA, C/. Pau Claris 97, 4º, 1ª Barcelona 08009, Spain*
c Kineo Mobility Analytics, Melior Business Centre - Diego de León 47, 28006 Madrid, Spain

Abstract

OD matrices that describe mobility patterns provide major input to most transport analysis models. Since OD matrices are not yet directly observable, they are usually estimated indirectly. Data from new sources such as mobile phone records and GPS traces from mobile apps are emerging alternatives that allow for cheaper and timely estimates. However, they are still hindered by weaknesses that must be studied for use as input to transportation models. This paper presents a case study using mobility data from mobile phone records, with the goal of establishing a methodology for validating the obtained OD matrices and generating the appropriate input for traffic assignment models. Spatial and temporal consistency of OD data elaborated from mobile records has been proved to be useful for transportation modelling needs.

Keywords: OD Estimation; Smartphone Data; Traffic Assignment

1. Introduction

Origin to Destination (OD) matrices are one of the main inputs to most transport analysis procedures and to all transport analysis models, whether they are based on static assignment, dynamic assignment, or traffic simulation. OD matrices describe mobility patterns in a selected geographical area, usually after it has been partitioned into a set of Traffic Analysis Zones that are conventionally identified as origins and destinations of trips across the area. Travel

* Corresponding author. Tel.: +349-3401-1038; fax: +349-3401-6575
E-mail address: lidia.montero@upc.edu.

10.1016/j.trpro.2018.12.211

demand analysis is the first step in the traditional 4 steps of Transport Models, and it is usually based on a well-established methodology that consists of detailed travel household surveys that are accurately designed, supported by careful sampling procedures, and whose results are combined with socioeconomic census data. Unfortunately, these procedures have some drawbacks: they are very expensive, so they cannot be repeated as frequently as necessary; data processing is complex and usually takes a significant amount of time, which leads to the availability of results being delayed; and, finally, they provide a static picture of the mobility patterns only at the time of the survey. As a result of these drawbacks, time dependencies and variabilities cannot be taken into account.

The advent of Information and Communication Technologies is changing the OD estimation scenario. In particular, access to mobile phone data sets offers an unprecedented possibility to investigate mobility patterns. The type of accessible anonymized data can vary widely: from aggregated records representing the number of unique devices using a given antenna for each specific time interval (e.g., one hour or over the whole day); to the so-called Call Detail Records (CDR) using various formats (e.g., anonymized user ID, longitude, latitude or time stamp). Researchers have recently addressed the problem of how to extract OD mobility patterns from this data by means of data processing approaches that extract stays and pass-byes from raw data. In addition, they have derived procedures for identifying location types, which typically indicate home, work or other. By processing this first level data in combination with data from other sources (usually census data), the OD matrix can ultimately be extended from the sample of data records to the whole population. Since this is a hot research topic, there is a rich literature that reports on recent findings, some examples of which would be: (Friedrich et al., 2010), who provide a seminal analysis; (Çolak et al., 2015), who provide deeper and more mature insight; (Chen et al., 2016), who present a survey of the state-of-the art; and (Jiang et al., 2016), who describe a relevant operational framework. Files delivered by KINEO contain 444 millions of OD registers, each containing 13 fields. A high performance PC including a 50 Gb RAM memory has been used to transform KINEO's OD data to formats needed by statistical analysis using RStudio.

1.1. Research Contribution

However, the resulting matrices still need to go through additional checking and validation processes before they can be used for transport analysis, especially considering that they will eventually provide major input to transport models. A variety of checking and validation process have been proposed, although there nevertheless remain many alternatives yet to explore. The first objective of the research reported in this paper is to propose a methodological approach for systematically analysing the spatiotemporal structure of the trip patterns represented by the generated OD matrix. This methodological approach is based on using Principal Component Analysis and Correspondence Analysis to identify the main underlying structures and relationships, which in turn will then validate whether or not the OD matrix coincides with the expected or known structures of the case being analysed. In our case, our study area is the Primary Crown of the Metropolitan Area of Barcelona, the proposed methodological process has been applied to the OD matrix, which was provided by KINEO, using raw mobile data supplied by Orange Spain.

Even after having been validated and established as useful in many studies on population dynamics, these OD matrices still have a critical drawback that limits their use in transportation analysis. While the use of map matching techniques allows reliable identification of mode and route choice for medium- and long-distance travel (García et al., 2016), the spatio-temporal resolution of the mobile phone data makes it more difficult to apply this approach to short trips, especially in urban areas, where it is often impossible to accurately distinguish the transportation mode (i.e., walking, biking, passenger car, bus, metro, local railway, etc.). In order to split the global OD into a set of OD matrices per transportation mode available in the region under study, an additional step is still necessary: the fusion of the primary OD matrices with other data sources that contain suitable information. This step is the second objective of the research reported in this paper.

2. Methodology

Fig. 1 depicts a synthetic view of the proposed research framework that is used to generate the multimodal OD inputs for transport models. This paper reports on the methods and procedures used in Phase I of the framework in order to validate, refine and generate the multimodal target matrices that can later be adjusted in subsequent phases (not covered in this work). "Old ODs" in Additional Data box refers to (TMB, 2007).

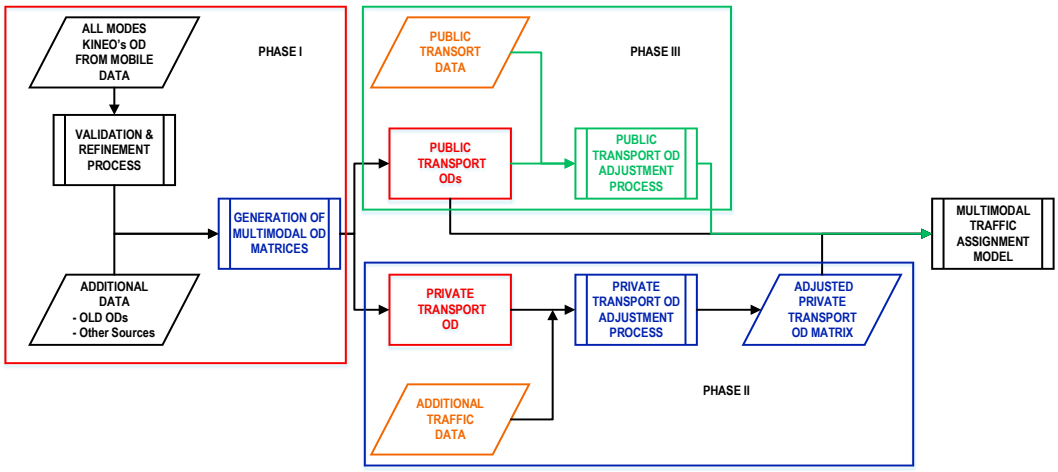


Fig. 1. Methodological framework for generating multimodal OD matrices from mobile phone data

The logic diagram in Fig. 1 illustrates the building blocks or groups of building blocks such as:

- Input data to KINEO's Data Processing Tools.
- Consistency Analysis (data post-processing).
- Modal Splitting Module.

The input data to KINEO's Data Processing Tools are namely three blocks:

- The sample of raw data from mobile phones for the "Call Detail Records" (CDR) supplied by Orange to KINEO for residents over 16 years old in the Metropolitan Region of Barcelona (AMB).
- The specification of the geographic scenario to be covered by the transport model (in this case, the Primary Crown of AMB) and the shapefile of the Transport Analysis Zones (TAZ) covering the geographic scenario.
- The census data. Census tracts from 2015 are composed of 1908 units for the whole area. Previous studies (EMIT, 2007) used a Zoning System aggregated to 131 for the whole of Catalonia and around 120 TAZ for the Primary Crown.

A TAZ that is well balanced in terms of population and demand analysis criteria is produced. This is a clustering exercise that covers the geographic scenario by using the census tracts. The methodological steps for obtaining the Barcelona Virtual Mobility Lab Zoning System (VML-TAZ) are:

- Load 2015 Census Tract shapefiles, including area and 2016 population, into QGIS and VISUM.
- Load 2007 EMIT shapefiles into VISUM/QGIS. Combine EMIT-TAZ units and 2015 Census Tracts to define EMIT-TAZ mapping in the 2015 Census Tracts. Define percentages for each 2015 Census Tract in EMIT-TAZ using QGIS tools.
- EMIT-TAZ population in 2016 must be estimated based on mapping the 2015 Census Tracts onto the EMIT-TAZ zones. Overall, the mean EMIT-TAZ population is over 20,000 inhabitants, but it is not uniformly distributed across the study area, which leads to some dramatic EMIT-TAZ imbalances such as Sant Boi Ll, which has 82,000 inhabitants in one TAZ on the border of the Primary Crown. Export to excel and MatLab the EMIT-TAZ composition, specifically in terms of the 2015 Census Tracts.
- The 2007 Modal OD matrices in the EMIT-TAZ zoning system are loaded into MatLab for future use together with the corresponding mapping of the 2015 Census Tracts.
- The VML-TAZ system is defined via complex clustering, which itself must be done manually by means of assessing the most convenient grouping. This exercise relies on spreadsheet support and GIS overlapping, as well as displays of the 2015 Section Tracts, EMIT-TAZs, municipality limits and district limits. The VML-TAZ zoning criteria are:

  - Each VML-TAZ unit should contain the 2015 Census Tracts pertaining to the same district and municipality. Cross-district and cross-municipality TAZ are not allowed.
  - VML-TAZ should be homogeneous in terms of land use, transport infrastructure access and socio-economical level.
  - The average 2016 population for each VML-TAZ inside every district should be 5,000 inhabitants, and the coefficient of variation across the district should be less than 0.5 for 80% of the municipal districts.
  - Harbour, airport, industrial polygons and large commercial areas have specific VML-TAZs.
  - VML-TAZ should avoid zones with long edges. Squared or circular shapes are preferred in order to thus enforce regularity.
  - For the sake of comparability with previous studies (EMIT), some compatibility with the EMIT-TAZ system should be preserved. EMIT-TAZ was defined using 2005 Section Tracts, which are not compatible with the 2015 Section Tracts; therefore, approximate EMIT-TAZ to VML-TAZ matching has been calculated using GIS tools.
- A cellular cell should not fully contain several VML-TAZ in low density areas, or it would otherwise not be possible to assign the proper origin. Additionally, it is not convenient to define very small VML-TAZ in centralized, high density urban areas, because this complicates the process of using CDRs. VML-TAZ shapefiles were discussed with KINEO.
- Hourly OD matrices that were representative of all modes of OD trips were obtained after processing 3 days of data from March 2017. OD matrices were segmented by gender, age group (4 classes, 16-24, 25-44, 45-64 and 65-99), purpose (4 classes: Home, Work, Other Frequent and Other Unfrequently) and residence area of trip-maker in three groups (Barcelona, First Crown and Metropolitan Region).

3. Research Contribution for Phase 1

The consistency analysis shown in Fig. 1 consists of a traditional mobility analysis that takes into account the entire work day in aggregated subareas. These global figures are also available from other sources (TMB, 2007; EMEF, 2015). A first step in establishing the quality of KINEO's data could be checking that there are no significant discrepancies among the global figures found in the available official sources. Internal trips in each subarea and the commuters moving between them account for the total amount of trips by all modes in the core area (Primary Crown) and the rest of the Metropolitan Region of Barcelona (an area that includes the Primary Crown). By fusing the data on the modal split that has been gathered from available sources (EMEF 2015), it is possible to estimate the modal split between macrozones in the Metropolitan Area. The modal splitting process is mandatory, since modal matrices are needed to feed a multimodal model of the Primary Crown (Montero et al., 2018). Finally, a new systematic analytics-based procedure has been applied to validate spatial OD patterns. Principal Component Analysis and Correspondence Analysis are in fact not new methods; but to the best of our knowledge, they have never been applied to OD pattern validation.

4. Consistency Analysis

The consistency of the 2017 OD matrices was demonstrated by assessing discrepancies and variations from previous studies based on global figures. Comparison to (EMEF, 2015) was given priority, since it contains the most updated data. (EMEF, 2015) considers the AMB area (Àrea Metropolitana de Barcelona, a larger study area), but some results for the Primary Crown also appear, such as:

- EMEF Total daily trips for the Primary Crown are about 9,100,000 trips, while KINEO figures are about 8,700,000 trips (<8% decrement).
- EMEF Total daily trips for AMB are about 11,700,000 trips, while KINEO figures for RMB (not AMB) are 14,851,415 trips.
- EMEF Total daily trips inside Barcelona are about 4,990,000, while KINEO figures are 4,795,769 trips.

Consistent values are delivered from calculating self-containment (the proportion of internal trips to one area out of the total number of trips in that area for modes) in Barcelona's Primary Crown and the whole Metropolitan Region

(EMEF 2015). According to EMEF, it is 0.88 (taking all trips generated within and attracted to the city by considering trip makers living in Barcelona), but it is 0.55 according to the 2017 data (taking all trips generated by and attracted to the city without restricting the resident area of trip makers). Total OD trips (all modes) in 2017 ranges between 0 and 259 trips in 99% of OD pairs, where 17.5% contain 0 trips and 29% of OD pairs contain a maximum of 1 trip. Median is 4 trips per OD pair, but when zeros are removed the central trend rises to 7 trips. A clean exponential distributed profile is shown for OD pairs involved with the Primary Crown. Once EMIT daily matrices have been expanded to the VML-TAZ system, both show around 20% of zero values. When the matrices are estimated from surveys, the percentage of 0 cells lies between 80-85%.

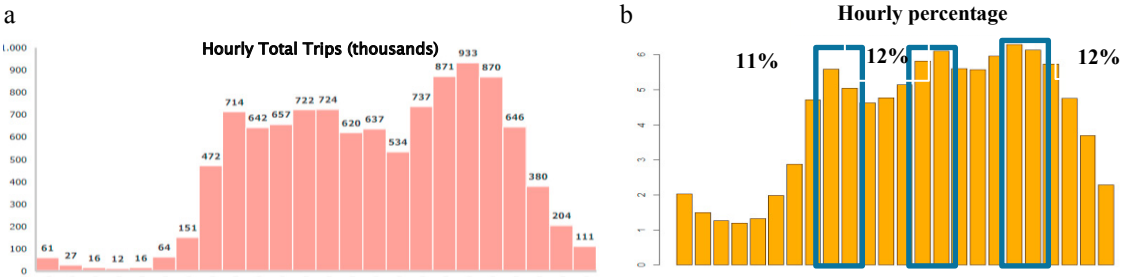


Fig. 2. Hourly trip profiles produced by mobile phone data (all modes): internal trips in the Primary Crown. (a) EMEF 2015 data; (b) 2017 KINEO data.

As a second step, the hourly pattern of trips produced over the whole day allows identification of the morning and afternoon peaks in the study area so that they can be contrasted with the profile per hour that is based on traditional loop detector data and indicated in reports for mobility authorities. This consistency analysis is complemented by taking the hourly pattern of trips produced according to mobile phone data (see Fig. 2 b) and comparing it to the (EMEF, 2015) pattern. A midday peak is highlighted in the 2017 data. Short trips are usually not reported in traditional travel surveys, and this might correspond to lunchtime mobility that is usually not motorized. The third step consisted of refining the two first steps for the target municipalities in the Primary Crown of Barcelona city and the rest of the Primary Crown. Depending on the availability of data for validation purposes, self-containment can also be considered at district levels in Barcelona and/or municipalities in the Primary Crown.

5. Modal Splitting Module

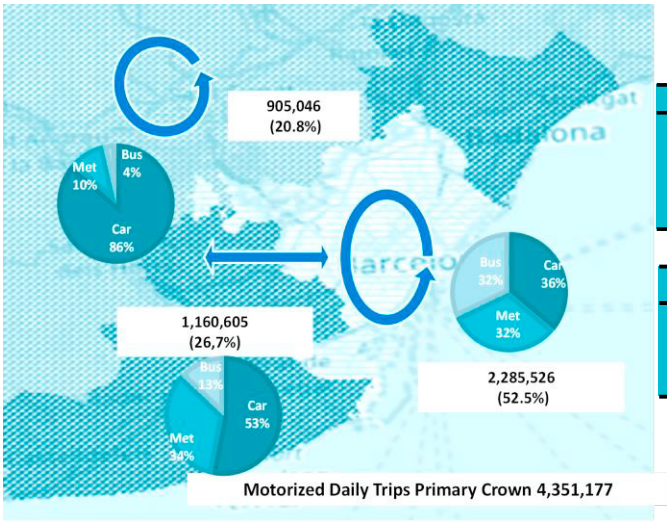
Elaborated data from CDR mobile phone records cannot identify modal splits on short trips, such as those that are common in urban areas. The developed procedure considers the aggregated modal split available at the EMIT-TAZ system level and it has been combined with the total trips (all modes) of the 2017 OD data at the VML-TAZ system level. The procedure can be summarized in the following steps:

- Let T_{IJ} be the total number of trips between an I-J pair of macrozones (EMIT-TAZ), and let $t(i_k,j_l)$ be the total number of trips between VML-TAZ i_k-j_l, where EMIT-TAZ I is composed of VML-TAZ $\{i_1, \ldots, i_K\}$, and EMIT-TAZ J is composed of VML-TAZ $\{j_1, \ldots, j_L\}$; thus, adding up $t(i_k,j_l)$ for all k and j, T_{IJ} is obtained, although each $t(i_k,j_l)$ is not T_{IJ}/KxL but is instead proportional to the VML-TAZ i_k and j_l populations. Let P_I and P_J be the total population of EMIT-TAZ I and J, respectively, and let Pi_k and Pj_l be the total population of the i_k and j_l VML-TAZs.

$$t(i_k, j_l) = P_{i_k} P_{j_l} T_{IJ} / P_I P_J \quad (1)$$

- If the previous procedure is applied to each modal EMIT-TAZ system matrix (car-bus-met and other), at the end, $t^m(i_k, j_l)$ are obtained for the m modes of car, bus, met and other. We have imposed the restriction $t^m(i,i) = 0$ for the modes of car, bus and met. These are modal EMIT-TAZ matrices converted to the VML-TAZ system.

- The previous procedure was implemented in MatLab (The MathWorks, Inc., Natick, Massachusetts 2016) to extend the modal split in macroareas to TAZ in the VML zoning system and has been applied to all-modal matrices of the VML-TAZ zoning System (homogeneously to periods).
- The EMIT modal matrices are expanded to the VML-TAZ zoning System and loaded into VISUM, and the OD modal split proportions are defined by calculating the formula matrices. In the future, these formula matrices can be applied to obtain all-modal matrix (VML-TAZ System). In particular, we have grouped the bus plus met modes to define the public and private transport OD matrices for 2017. Modal splitting has been applied homogenously to each hourly period, since no additional data were available at hour granularity.



Modal Split	bus	car	met	All
IntraPC	47,923	766,883	90,240	**905,046**
IntraBCN	728,523	814,767	742,236	**2,285,526**
Commuters	151,726	617,053	391,826	**1,160,605**
	928,172	**2,198,703**	**1,224,302**	**4,351,177**

Modal Split	bus	car	met	Subarea %
IntraPC	5.3%	84.7%	10.0%	**20.8%**
IntraBCN	31.9%	35.6%	32.5%	**52.5%**
Commuters	13.1%	53.2%	33.8%	**26.7%**

Fig. 3. Daily motorized mobility and modal split. Modes: bus, car and met (metro, train and tram). Row labels: IntraPC for Internal Trips to Primary Crown (excluding Barcelona), IntraBCN for internal trips to Barcelona and Commuters for trips from/to Barcelona.

At the aggregated level and after fusing with modal split data (TMB 2007), motorized total trips for Primary Crown are shown in Fig. 3.

6. Analytics Procedures for Mobility analysis

Normalized Principal Component Analysis (PCA) has been applied to the hourly 2017 OD total trips (internal Primary Crown – all OD pairs) to clarify the hourly matrix relationship. PCA reveals the hidden structural information in large sets of multidimensional data: X(n,p) (n and p large, np very large). The active variables for PCA analysis are hourly OD trips (24 columns, 624x624 rows). Furthermore, once the factorial axes have been calculated, the supplementary variables for projection and interpretation are the total daily trips for 2017and 2008, according to the source EMIT, and daily 2017 total OD trips. OD matrices are considered in tabular form, so they appear as a column (variable) in PCA analysis. The first finding is that hourly OD trips seem to be highly correlated among each other (0.85 to 0.89) (see

Fig. 4). After an in-depth analysis, we discovered outliers: less than 1% of OD pairs dominate the principal component calculations (a non-robust PCA procedure was used). Hourly matrices are seriously affected by short trips pertaining to pedestrian (or non-motorized) trips, which often occur as intra-district mobility. After selecting those OD pairs that have very large coordinates on the first factorial plane (those that can be considered multivariate outliers because they have greater than a 95% percentile), these trips belong to pedestrian intrazonal trips and intra-Barcelona TAZs. Those observations are considered to be supplementary and non-active for factorial axis calculations, thus providing the plot in

Fig. 4. The first axis is a size axis, with positive values being related to OD pairs for Barcelona and negative values to those which are non-Barcelona. The first axis explains 25% of the variability in data and the second axis 5%. Only third and fourth factorial axes are relevant, according to Kayser's rule (eigenvalues greater than 1), thus leaving a lot of variability to explain, which is consistent with demand analysis. Hourly patterns that correlate positively with the second axis are morning patterns and while those that correlate negatively are afternoon patterns. As a consequence, hourly patterns seem to be consistent in the sense that morning-hour variables are plotted closer to each other (with a

sharp angle between arrows) and opposite to the afternoon hours. Thus, the temporal granularity that it is very useful for dynamic models can be obtained from mobile phone data; and this is a highly important finding.

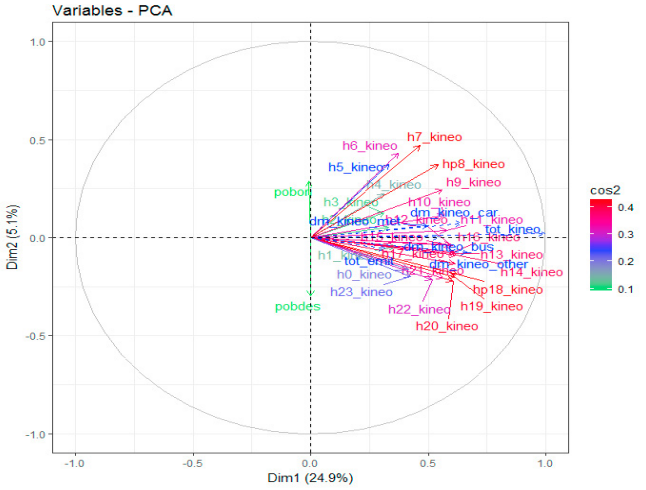


Fig. 4. PCA Analysis of OD trips per hour with all OD pairs (outliers removed). Colours refer to squared cosines between variables and the factorial axis.

The spatial consistency of OD matrices is laborious to check in practice. Desire line plots (the most important origin-destination relationships) are usually plotted and interpreted according to knowledge of the study area, and they are commonly available in professional modelling platforms such as VISUM. Contingency tables between municipalities for each mode (or all modes together) can be easily calculated either with professional modelling platforms or by loading tabular OD modal matrices into Excel, aggregating TAZ with dynamic tables that map the VML-TAZ zone onto districts and municipalities. However, marginal aggregated tables with percentages by rows or columns are usually needed.

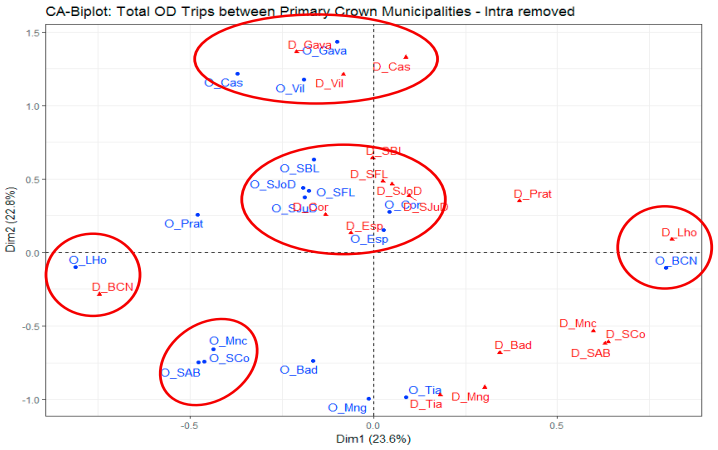


Fig. 5. Correspondence Analysis projection of Total OD daily trips between municipalities in the Primary Crown.

The main analysis is based on contingency tables, and analytics provides tools for analyzing large contingency tables by using graphical tools that are supported by rigorous factorial analysis, such as Correspondence Analysis (CA). We have applied CA to validate spatial consistency of OD flows since singular OD relations are highlighted. We use R (R Development Core Team 2016) in order to analyse:

- Contingency tables of total and modal trips between municipalities in the Primary Crown (18x18 matrices),
- Contingency tables of total and modal inter-district trips between Barcelona districts (10x10 matrices).

- Contingency tables of total and modal trips between Barcelona inter-neighbourhoods (73x73 matrices).

The Biplot in Fig. 5 is the first factorial plane projection showing the total generated trips (O for origin) and total attracted trips (D for destination) at the level of municipal aggregation. Points closer to each other indicate a common pattern. For example, O_Gava and D_Gava are very close (blue for trips generated at Gavà and red for trips attracted to Gavà), meaning that most of the generated trips at Gavà are destined for the same municipality of Gavà. Most trips generated at L'Hospitalet (the second-most populous city in Catalonia) are remarkably attracted to Barcelona. However, this behaviour is not shared by other surrounding large municipalities such as Badalona. CA biplots avoids percentage calculations and ranks the spatial dependency between subareas.

7. Conclusions and Further Research

The 2017 OD matrices that were elaborated for modelling purposes from mobile phone CDRs are available in a detailed zoning system (VML-TAZ) for the Primary Crown of Barcelona. The hourly OD matrices for 2017 have been successfully checked for consistency against the available official sources. A modal splitting procedure has been applied to obtain modal matrices, and it has also been successfully validated. An innovative analysis of temporal and spatial relationships has been applied. Although PCA and CA are not new methods, they have not been applied previously to demand analysis, and they are very promising to understand mobility patterns in large areas.

Acknowledgements

This research was funded by the TRA2016-76914-C3-1-P Spanish R+D Program and by Secretaria d'Universitats-i- Recerca -Generalitat de Catalunya- 2017-SGR-1749.

References

Chen, Cynthia, Jingtao Ma, Yusak Susilo, Yu Liu, and Menglin Wang. 2016. "The Promises of Big Data and Small Data for Travel Behavior (Aka Human Mobility) Analysis." *Transportation Research Part C: Emerging Technologies*. https://doi.org/10.1016/j.trc.2016.04.005.

Çolak, Serdar, Lauren P Alexander, Bernardo G Alvim, Shomik R Mehndiratta, and Marta C González. 2015. "Analyzing Cell Phone Location Data for Urban Travel." *Transportation Research Record: Journal of the Transportation Research Board* 2526 (January): 126–35. https://doi.org/10.3141/2526-14.

EMEF. 2015. "Daily Mobility Data in Metropolitan Area of Barcelona (Enquesta de Mobilitat En Dia Feiner-EMEF)." Enquesta de Mobilitat En Dia Feiner (EMEF), ATM, AMB Aj. de Barcelona, AMTU i IDESCAT. 2015. https://www.amtu.cat/enquestes-de-mobilitat-interurbana/1746-emef-2015.

Friedrich, Markus, Katrin Immisch, Prokop Jehlicka, Thomas Otterstätter, and Johannes Schlaich. 2010. "Generating Origin-Destination Matrices from Mobile Phone Trajectories." *Transportation Research Record: Journal of the Transportation Research Board* 2196 (December): 93–101. https://doi.org/10.3141/2196-10.

García, Pedro, Ricardo Herranz, José Javier Ramasco, Gennady Andrienko, and Nicole Adler. 2016. "Big Data Analytics for a Passenger-Centric Air Traffic Management System A Case Study of Door-to-Door Intermodal Passenger Journey Inferred from Mobile Phone Data." In *Proceedings of the 6th SESAR Innovation Days*. Delft: D. Schaefer Editor.

Jiang, Shan, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shounak Athavale, and Marta C González. 2016. "The TimeGeo Modeling Framework for Urban Motility without Travel Surveys." *Proceedings of the National Academy of Sciences of the United States of America* 113 (37): E5370-8. https://doi.org/10.1073/pnas.1524261113.

Montero, L., M.P. Linares, J. Salmerón, G. Recio, E. Lorente, and J.J. Vázquez. 2018. "Barcelona Virtual Mobility Lab: A Multimodal Transport Simulation Testbed for Emerging Mobility Concepts Evaluation." In *Proceeding of the Ninth International Conference on Cloud Computing, GRIDs, and Virtualization - CLOUD COMPUTING 2018 February 18, 2018 to February 22, 2018 - Barcelona, Spain*. https://www.iaria.org/conferences2018/CLOUDCOMPUTING18.html.

R Development Core Team. 2016. "R: A Language and Environment for Statistical Computing." *R Foundation for Statistical Computing*. https://doi.org/10.1007/978-3-540-74686-7.

The MathWorks, Inc., Natick, Massachusetts, United States. 2016. "MATLAB and Statistics Toolbox Release 2016a."

TMB. 2007. "EMIT - Enquesta de Mobilitat a La RMB de Transport Metropolitans de Barcelona (TMB)." https://transportpublic.org/images/pdf/20081000-estudi_tmb.pdf.