Graells-Garrido et al. *EPJ Data Science* (2018) 7:49
https://doi.org/10.1140/epjds/s13688-018-0177-1

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Inferring modes of transportation using mobile phone data

Eduardo Graells-Garrido[1,2*], Diego Caro[1,2] and Denis Parra[3]

*Correspondence: egraells@udd.cl
[1]Data Science Institute, Faculty of Engineering, Universidad del Desarrollo, Santiago, Chile
[2]Telefonica R&D, Santiago, Chile
Full list of author information is available at the end of the article

Abstract

Cities are growing at a fast rate, and transportation networks need to adapt accordingly. To design, plan, and manage transportation networks, domain experts need data that reflect how people move from one place to another, at what times, for what purpose, and in what mode(s) of transportation. However, traditional data collection methods are not cost-effective or timely. For instance, travel surveys are very expensive, collected every ten years, a period of time that does not cope with quick city changes, and using a relatively small sample of people. In this paper, we propose an algorithmic pipeline to infer the distribution of mode of transportation usage in a city, using mobile phone network data. Our pipeline is based on a Topic-Supervised Non-Negative Matrix Factorization model, using a Weak-Labeling strategy on user trajectories with data obtained from open datasets, such as GTFS and OpenStreetMap. As a case study, we show results for the city of Santiago, Chile, which has a sophisticated intermodal public transportation system. Importantly, our pipeline delivers coherent results that are explainable, with interpretable parameters at each step. Finally, we discuss the potential applications and implications of such a system in transportation and urban planning.

Keywords: Mobile phone networks; Urban informatics; Commuting; Non-negative matrix factorization; Mode of transportation

1 Introduction

People spend their time not only *within places*, but also *moving from one place to another*. As Charles Montgomery says in his book, Happy City: *"City life is as much about moving* through *landscapes as it is about being* in *them"* [1]. Some trips are crucial in people's lives, such as the trip from home to work, and *vice versa*. This recurrent activity, called *commuting*, has several effects in quality of life, both, positive and negative [2, 3]. For instance, for some people it is the least liked daily activity [4]. Therefore, an understanding of commuting patterns would provide opportunities to improve quality of life at scale in a city. Moreover, by understanding commuting, it would be possible to inform public-policy design, the planning of transportation networks, and correlate commuting to factors such as health, social habits, exposure to pollution, stress, among others.

Traditionally, commuting has been studied with well-known methods such as travel surveys [5], focus groups [6], and traffic counts [7]. Surveys have a number of drawbacks, including the lack of repeated observations over time and reporting biases and errors [8, 9]; focus groups may allow for repeated observations, but their sample size tends to be

very small; and traffic counts are not scalable at the city level. In modern contexts, these methods are not able to keep the pace of city growth and change, making relevant dynamic phenomena to be invisible for transportation and urban planners.

The availability of large amounts of digital traces has allowed to study urban phenomena at spatio-temporal granularities that traditional methods cannot. One of these data sources is the set of billing records from mobile phone networks, known as *Data Detail Records* (XDR) [10]. XDR provides a cost-effective way to perform studies about human behavior [11], because mobile operators already generate, store, and analyze the data for billing and marketing purposes.

In this paper, we seek to answer the following research question: *how to infer the distribution of mode(s) of transportation in commuting within a city using mobile phone network data?* The answer would provide insights to manage, plan, and design urban transportation systems, urban infrastructure, and public-policy, among other applications. To do so, we propose an interdisciplinary approach: using methods and tools from Data Science [12] and Transportation [13], we define a pipeline that is able to infer one or two modes of transportation chosen for commuting by users.

The main step of our pipeline refers to the analysis of trips inferred from XDR. We focus on the billing records generated *while* commuting, which we represent in a *waypoint matrix*, similar to document-term matrices in Information Retrieval [14]. We decompose this matrix using Topic-Supervised Non-Negative Matrix Factorization (TS-NMF), a method that mixes NMF [15] with Semi-supervision [16] through Weak-Labeling [17]. The pipeline extends our prior work: in [18] we proposed a method to infer trips from XDR, and in [19] we explored how plain NMF behaved when decomposing *waypoint matrices*.

Here we present the following contributions: (i) A processing pipeline that, given XDR and auxiliary data commonly available, generates the distribution of mode(s) of transportation usage for commuting in a city; (ii) A case study that evaluates the proposed pipeline in a city with more than seven million inhabitants and a public transportation system designed for intermodality; and (iii) A discussion about the implications of using the proposed pipeline for transportation analysis, on the basis of its explainability.

2 Methods

In XDR, the main unit of analysis is a *network event* [20], which indicates a billing record for a device. Such events include calls, text and multimedia messages, and Internet downloads. Calls and messages are billed individually, Internet connections are billed in batches. The number of TCP/IP network packages sent through a tower may be very high, but billing is performed according to the number of megabytes transmitted. We work with anonymized XDR data, where each network event contains the ID of the tower, a timestamp, and a tokenized device ID. Device IDs are coherent in the dataset, *i.e.*, two events with the same ID describe the trajectory of the same device.

In transportation, the core unit of analysis is a *trip*, with its corresponding attributes, *e.g.*, trip origin, destination, departure time, traveled distance, purpose, and mode(s) of transportation [13]. Trips can be aggregated into Origin-Destination (OD) matrices, which encode the number of trips from one area of the city to another. These areas can be blocks, neighborhoods, municipalities, among other administrative divisions. Transportation experts usually work with OD matrices that are representative at the county or municipal

level, due to the limitations of data-collection methods. For instance, in Santiago, Chile, the last travel survey is representative at the municipal-level, meaning that, even though there is individual trip information available, only the municipal-level analysis is representative of city behavior.

This paper focuses on a transportation problem: *the inference of mode(s) of transportation for commuting*. Since commuting refers to a recurrent, routinary trip, it is possible to go one level up in the analysis pipeline, and move from *trips* to *commuters*. Thus, our main unit of analysis will be devices used by commuters. From now on, we refer to commuters or users indistinctively.

To solve the problem, we propose a pipeline that takes XDR as input, generates a list of commuters, with home and work location, and their assigned mode(s) of transportation. The result can be a single mode or a combination (*e.g.*, bus and metro). Figure 1 shows a schematic view of the problem. Starting from XDR, an algorithm infers trips for each device (Section 2.1). These trips are then used to identify home and work locations for a device, allowing to assign trip purpose, and thus, effectively labeling commuting trips (Section 2.2). Next, each tower is labeled according to their proximity to relevant urban/transportation infrastructure, using crowd-sourced geographical data and public transportation network feeds (Section 2.3). With the set of non-pedestrian commuting trips, a user-tower matrix is built, according to the towers that users connected to while moving. Using the tower labels, some users that do not perform pedestrian trips can be weakly-associated to a mode of transportation. These users are considered as seeds for a semi-supervised model (Section 2.4), which, is next used to group users into *modal clusters* (Section 2.5). In those cases that do not have enough information as input for the model, we identify whether the unlabeled home/work trajectory is pedestrian (Sec-

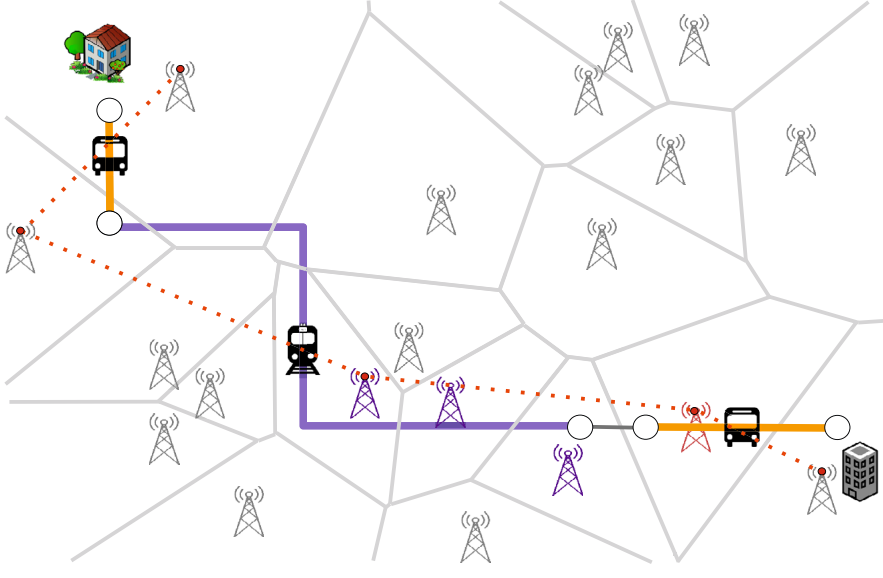

Figure 1 Example of a XDR trajectory for a trip from home to work. The trip is intermodal, with several stages: bus, metro, and bus. The red dotted line depicts the trajectory built by following the user connections. The assignment of towers to each XDR event is approximated by a Voronoi tessellation of the space according to the tower positions. Some mobile phone towers are close to the route followed by the user, however, the user does not necessarily connect to all of them (*e.g.* his/her device could connect to towers far from the route)

tion 2.6). Finally, by aggregating all commuters and their labels, we are able to estimate the distribution of usage of mode(s) of transportation in a city, also known as *modal partition* in transportation terms (Section 2.7).

The remaining part of this section explains each stage of the pipeline in detail.

2.1 Trip inference through activity detection

This stage models the task of inferring trips for a given spatio-temporal trajectory using computational geometry techniques and transportation rules. A trip is considered one of the many activities that can be performed within a day. Let J be the set of tuples for device u at a given day:

$$J_u = \left\{ \left(A_i, (t_{iO}, t_{iD}), (p_{iO}, p_{iD}), I_i \right) \right\}. \tag{1}$$

In the tuple i, A_i is an activity type, p_i (and t_i) are the positions (and times) associated to the origin (p_{iO}) and destination (p_{iD}) of A_i, and I_i is the set of intermediary points in the trajectory from p_{iO} to p_{iD}. Activities can be of types *trip*, *stay*, and *unknown* (*i.e.*, activities that cannot be classified due to lack of data). The intermediary points are denoted as *within-trip waypoints*.

To identify activities, we assume that, during a day, a set of turning points exist [18]. A turning point is a moment of the day, at a specific position, where the device owner started to perform an activity (and, by definition, ends performing a previous activity). To build the list of turning points, we define the following vector per each user u:

$$\vec{u} = \left[(t_0, p_0), (t_1, p_1), \ldots, (t_n, p_n) \right], \tag{2}$$

where each element in $\vec{u}$ corresponds to an event of u in a day, with a timestamp t, and a tower position p. These vectors can be projected into a 2D plane: the x-axis is the elapsed time during the day, and the y-axis is the accumulated distance from the starting point of the day:

$$d_i = \sum_{j=1}^{i} E(p_j, p_{j-1}), \tag{3}$$

where E is the Euclidean distance function between two points in space.

Next, we build a spatio-temporal trajectory over the turning points of $\vec{u}$:

$$T_u = \left[(t_i, d_i) : \forall i \in [0, n] \right]. \tag{4}$$

To identify turning points, we simplify T_u into S_u using the Visvalingam–Whyatt line simplification algorithm [21]: the points identified as relevant by the algorithm are considered turning points. The algorithm assigns a weight to each point in the trajectory, and keeps only those with a weight above a given threshold. The weight of a point is defined as the Euclidean area formed by the triangle of the previous, current, and following point. The greater the weight, the greater the importance of the point. By definition, the starting and ending points of the trajectory have infinite weight, and thus, under any threshold they are always present in the result of the algorithm. Note that prior work [18] used a different

algorithm (Ramer–Douglas–Pecker), however, Visvalingam–Whyatt has a threshold with interpretable units, namely, distance multiplied by time. The points from T_u not included in S_u are saved into I_i as waypoints.

The points in S_u are chained to build a list of segments that represent activities. To do so, we employ a set of rules. *Unknown* segments are those where the total covered distance is greater than 50 kilometers. In those cases we cannot distinguish between trips and unknown situations, such as when mobile phones are connected to distant towers that are on top of a hill, or when the mobile number is associated to a vehicle (*e.g.*, a taxi). While these are indeed displacements in time and space, they are so large for the city scale that the dataset may have missed inter-events (*e.g.*, due to connection to WiFi networks). *Stays* are stationary activities. Some stationary activities involve displacements (*e.g.*, working/studying in a big campus), but the speed of movement is much slower than when performing a trip. Thus, if there is a distance displacement, but the time is greater than 180 minutes, we still identify the activity as a stay. Finally, *trips* are segments that are not *unknown* or *stay*.

We merge contiguous segments tagged with the same activity. Two or more segments are merged into an activity by keeping the first time and position in the segment as origin, and the last time and position in the segment as destination. Additionally, there is an special case when two *trips* surround a *stay*. If the duration of the latter is lesser or equal than 15 minutes, its activity is changed to *trip*, and merged accordingly. This scenario corresponds to situations when users in public transport make a connection, or when vehicles are stuck in traffic.

After merging all activities, for each user there is daily set of activities J_u for each day in the dataset. Note that the turning points of merged activities are saved as waypoints in I_i.

2.2 Trip purpose

A commuting activity is a *trip* within two *stays*: one at home and one at work. This implies that, for a given device, we need to infer these two important locations: home and work.

In general, people follow daily routines where they spend most of the nights at home, and most of the hourly days at work in business days. This enables to infer these important locations in several ways, such as heuristics [22] or pattern recognition [23]. Given that we seek for interpretability in all stages of the pipeline, we implemented the heuristics defined in [22]: home is the most frequent area with stays at night, and work is a frequent area with stays at work hours that is more distant than others. This procedure allows to add an additional label to each trip in a set of activities J_u: whether it is a commuting trip or not.

2.3 Tower labeling using urban infrastructure data

In parallel to trip inference, we associate towers to modes of transportation as a way to provide weak labels to the inference process.

A tower provides connectivity to the devices around it, however, those devices may be within different contexts. For instance, the people connected to a tower installed within a metro station are more likely to be commuting than the people connected to a tower within a park. In a similar way, the people connected to a tower near a highway are more likely to be commuting by car than people connected to a tower in a main street, where several bus services are available. Likewise, car drivers are more likely to be on local streets

than in main ones [24], and passenger loads imply that, even if there are more cars than buses in main streets, a device in a main street with bus routes is arguably more likely to be in a bus than in a car. Regarding bikes, we leave them out of analysis. We explain the reasons in the Future Work section.

Having these assumptions into account, we associate each tower to one or more modes of transportation according to their proximity to urban infrastructure: highways and secondary streets are associated to cars; primary streets, to buses, if there are available bus routes; and bus corridors, to buses. Towers near metro over surface are also associated to metro. This last distinction is relevant as the underground metro network has dedicated towers identified as such. As result, for each mode of transportation m, we have a set of towers T_m that contains the corresponding associated towers. To increase distinction between several modes of transportation, we filter each T_m by removing towers that belong to more than one set.

2.4 User modeling from trajectories

Given the granularity of XDR, identifying the mode(s) of transportation of a single trip is unlikely. Consider a trip that lasts 45 minutes: according to a typical granularity of 15 minutes between records, in the best scenario this trip has three events: a trip start, a *within-trip waypoint*, and a trip end. Hence, to identify the mode(s) of transportation, we have only one event: the *within-trip waypoint*. To avoid this limitation we propose to aggregate commuting trips, which, by being recurrent, allow to have a complete picture of what is the urban infrastructure associated to user routines.

The first step is to build a *waypoint matrix* W, defined as:

$$w_{i,j} = \frac{\#\text{ of } \textit{within-trip} \text{ events of user } u_i \text{ at tower } t_j}{\#\text{ of } \textit{within-trip} \text{ events of user } u_i}. \tag{5}$$

This schema is equivalent to the row-wise normalized document-term matrices found in Information Retrieval [14], where users are the equivalent of documents, and towers are the equivalent of terms.

Our hypothesis is that, by decomposing W with matrix factorization, we will effectively arrange towers into clusters (or latent components) according to their co-occurrence in users' daily routines. To do so, we decompose this matrix into two:

$$W = A \times B, \tag{6}$$

where A is a $|u| \times k$ matrix that encodes k user latent features for $|u|$ users, and B is a $k \times |t|$ matrix that encodes k latent tower features for $|t|$ different cell towers. As matrix decomposition method we work with Non-Negative Matrix Factorization (NMF) [25, 26], in which by definition all $w_{i,j} \geq 0$. We choose NMF over other matrix decomposition methods such as SVD [27] (usually used to perform Principal Component Analysis or PCA [28]) because it has shown superior performance for the task of clustering [19, 29–31]. Moreover, the non-negativity constraint results in more interpretable latent features, since any user (rows) or tower (columns) in the W matrix can be represented as a weighted sum of parts [25, 29], all positive or zero. Then, using NMF, the matrix W is decomposed into two non-negative matrices, which gives a lower rank approximation for W, such that $W \approx A \times B$ [32]. For solving NMF, the problem has been formulated in

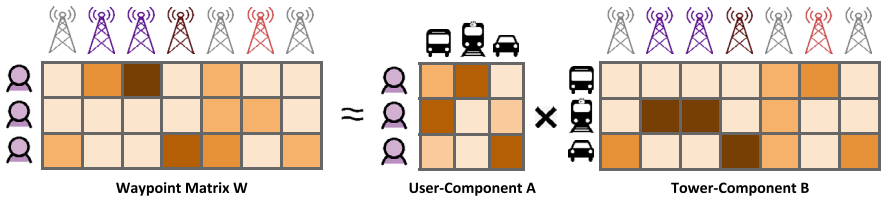

Figure 2 Depiction of the Non-Negative Matrix Factorization (NMF) stage of the proposed pipeline. We built a *Waypoint Matrix W* where rows encode users and columns encode towers, in analogy to document-term matrices. A cell in *W* denotes how frequently its corresponding user connects to its corresponding tower while commuting. The NMF decomposes *W* into two matrices that can be interpreted as the association between users and latent components (*A*), and towers and latent components (*B*). To interpret these latent components as modes of transportation, we used a Semi-Supervised approach to perform the NMF

several ways (*e.g.*, Frobenius and Kullback-Leibler losses [25, 31]), and different methods have been proposed to solve it (*e.g.*, multiplicative method, coordinate descent, *etc.* [33]). In our work, we start with the traditional formulation based on the Frobenius norm in order to further extend it to incorporate constraints based on our data. Figure 2 shows a diagram that explains the rationale behind using NMF to cluster users into modes of transportation with NMF.

NMF can be formalized as the following optimization problem:

$$\min_{A,B} \|W - A \times B\|_F, \tag{7}$$

subject to A and B be non-negative, where number of rows in A and the number of columns in B correspond to the desired lower-rank approximation k. In the original algorithm, the parameter k must be chosen manually, and its value should be decided jointly between data scientists and domain experts. In prior work [19], we found that, for several values of k, the clusters determined by NMF were of two types: urban areas delimited by contiguity, and transportation networks. As such, there is potential on guiding the algorithm to focus only in transportation features. To encode this prior information in the NMF algorithm, we use a semi-supervised approach named Topic-Supervised NMF (TS-NMF) [16]. With this method, we are able to provide examples to the algorithm about some users that we already know to which cluster they belong. This information is based on how we associated towers to modes of transportation in the previous step. Thus, we propose to use $k = 3$, where the clusters are *metro*, *bus*, and *car*. Based on MacMillan *et al.* [16], the previous formulation now incorporates constraints of users and modes of transports in a matrix L, and thus the original NMF formulation extends to:

$$\min_{A,B} \left\| W - (L \odot A) \times B \right\|_F^2, \tag{8}$$

where $\odot$ is the Hadamard product operator. The matrix L contains the user labels, defined as:

$$L_{u,m} = \begin{cases} 1, & \text{if } P(u|m) \geq h, \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

where h is a threshold probability (*e.g.*, 0.8). The factor $P(u|m)$ is the probability of user u being strongly associated to a specific mode of transportation m, given the set of towers T_m, defined as:

$$P(u|m) = \sum_{t \in T_m} w_{u,t}. \tag{10}$$

If other prior information is known, such as socio-economic information (*e.g.*, census data) of the area in which user u lives (inferred in a previous step), then $P(u|m)$ can be updated, for instance, using Bayes Theorem.

2.5 Inference of mode(s) of transportation

In this step we work with the matrix A, which contains the user associations with latent components. We first normalize the matrix row-wise, to convert it into a matrix A' of probabilities, such that $a_{u,m}$ is the inferred probability of user u choosing mode of transportation m for his/her commuting.

Since the several values of $a_{u,m}$ lie in the continuous range $[0, 1]$, we still need to assess the decision boundaries to classify u as user of specific modes of transportation. One way to interpret these values into a label such "metro" or "bus and metro" is by performing an additional clustering step on these associations. This step requires a manually specified parameter k', which should be higher than the number of latent components k from the previous stage; otherwise, intermodality will not be detected. Hence, we use k-means [34] to obtain *modal clusters*, which effectively quantize the rows of A'.

After quantization, a transportation expert may examine the centroids of each modal cluster, and then may proceed to label them, including his/her knowledge about the city into the model. For instance, a centroid $\{\text{metro} : 0.6, \text{bus} : 0.3, \text{car} : 0.1\}$ may be labeled either as "metro" or "metro and bus," depending on how the users closest to that centroid distribute in the city. Then, users are assigned a modal cluster label based on the expert's interpretation.

As result of this step, users have a tag that identifies their mode of transportation usage for commuting.

2.6 Identification of pedestrian trips

It is possible that some users do not generate *within-trip* events due to how they consume data from the Internet, or due to short trips that do not allow the billing cycle to capture events in the middle of a trip. In this step we try to classify those users that were not classified into specific modes of transportation into pedestrian commuters. Those users that were not classified into either are flagged with a null value.

Pedestrian trips have decision variables that differ from other modes of transportation, including distance, available infrastructure, and safety [35]. Of these factors, distance is arguably the most critical. As such, we label users as pedestrian or not based on their distance from home to work. This distance may be manually selected by knowing the typical walk distances in a city, through transportation studies [36] (which indicates a typical maximum of 750 meters for pedestrian trips), or fitted using regression if there is access to a labeled set of trips. In both cases, care must be taken due to the characteristics of mobile phone network data: trips have starting and destination *towers*, not specific locations.

2.7 Estimation of the modal partition

At this stage, for each commuter we have an assignment of a modal cluster, or a pedestrian flag, or a null flag. We discard those users without modal cluster or pedestrian flags, as we cannot classify them into a specific mode of transportation. Then, we aggregate users to estimate the *modal partition*, *i.e.*, the transportation term to denote the distribution of mode of transportation usage. Particularly, we follow two aggregation strategies: first, according to home locations; second, according to home/work location pairs into an Origin-Destination (OD) matrix. Here, location can be any administrative unit; surveys are typically representative at the county or municipal level. These aggregations are commonly used by transportation experts in their day to day work [13], and, with this pipeline, we expect to generate data that is coherent and comparable to those collected through surveys, with greater granularity levels—either spatial, temporal, or both.

3 Datasets and initial steps of the pipeline

In this section we describe the datasets employed in our case study, and the results up to the tower labeling stage of the pipeline. Our case study is performed in the urban area of Santiago, the capital of Chile. Santiago is a city with almost 8 million inhabitants, with 35 administrative units denoted *municipalities* within its urban area. For a description of its socio-economic characteristics in the same period of study, please refer to previous work using similar datasets [37, 38].

Santiago has a public transportation system named Transantiago [39]. By design, Transantiago provides feeder (bus) and trunk services (bus, metro), thus, given the extension of the city (837.89 km^2), it is expected that a relevant fraction of trips includes more than one mode of transportation.

We worked with the following datasets: (i) an anonymized XDR from Telefónica Movistar, the largest operator in Chile (30% market share), dated in August 2016; (ii) the socio-economic national survey CASEN held in 2015 in Chile, representative at the municipal level, which we use to build prior probabilities; (iii) an OpenStreetMap (OSM) dump of August 2016, and a General Transit Feed Specification (GTFS) of Transantiago for August 2016, which we use to associate towers to modes of transportation; and (iv) the travel survey held in 2012 in Santiago, Chile, representative at the municipal level, which we use to compare our results, and to define the areas of home and work locations in the pipeline.

3.1 CASEN survey

As prior information we used the CASEN (*CAracterización Socio-Económica*) survey. This survey is held every two years, and the 2015 edition is the last one released at the time of reporting this work. One of its questions is: *What is your choice of mode of transportation to go to work/study?* We used the answers to build a set of prior probabilities of using public transportation or cars (note that the answer "public transportation" does not specify bus, metro, or the potential usage of both).

This survey, including its expansion factor, considers a commuter population of 2,732,290 inhabitants in the municipalities under consideration. Figure 3 shows the modal partition per municipality, sorted by car usage. Figure 4 gives geographical context to this partition, and also depicts the segregation of the city through the distributions of public transportation (61.95% of trips) and car usage (22.25%). Note that non-motorized and

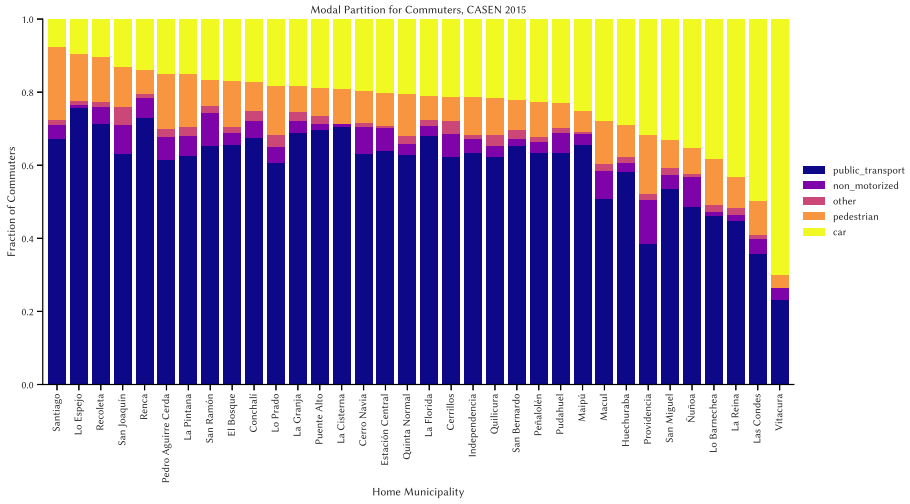

Figure 3 Relative distribution of mode of transportation usage (*modal partition*) per home municipality in Santiago, according to the CASEN survey. The municipalities are sorted with respect to car usage. The *public transport* category contains buses, metro, and the connections between both modes

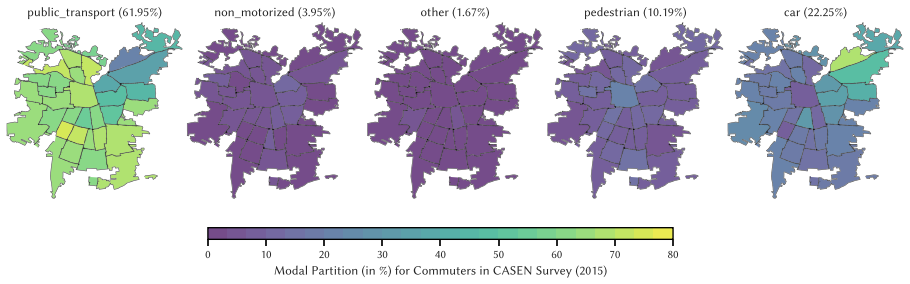

Figure 4 Spatial distribution of mode of transportation usage (*modal partition*) per home municipality in Santiago, according to the CASEN survey. The maps show how public transport and car usage reflect the socio-economic segregation of the city

other modes (*e.g.*, bikes) have a small share of the distribution. In addition to the lack of suitable infrastructure in the city that would allow to weakly-label users, this is one of the reasons we have not included these modes in our model. We discuss this further in the Future Work section.

3.2 Travel survey

To evaluate our model we used the Santiago Travel Survey held in 2012. It is the most recent travel survey available for the city (the previous one was from 2002). We considered the commuting trips of 15,116 respondents, who, after expansion with the survey weights, represent a population of 2,909,352 inhabitants. This number is similar to the expanded sample from the CASEN survey.

The travel survey defines a set of *Traffic Analysis Zones*, which have a finer spatial granularity than municipalities (there are 740 in the urban area under study). As unit of home and work locations we used these zones, depicted in Fig. 5. They are similar to census tracts, but have into account important factors for urban and transportation planning, such as floating population. Additionally, using zones allows to respect customer privacy,

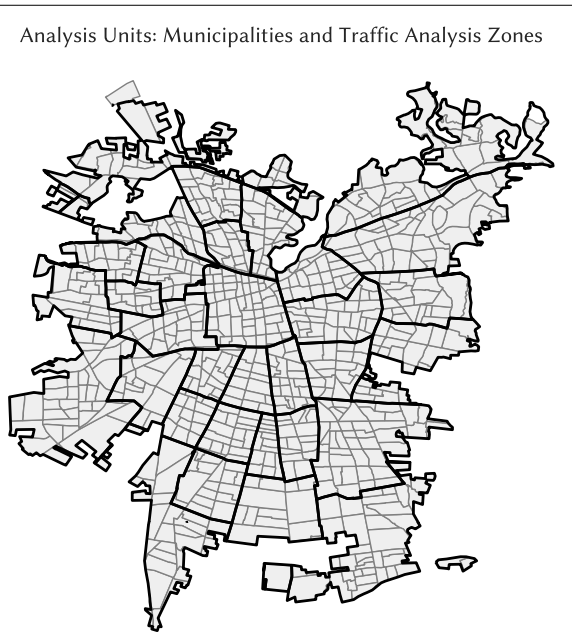

Figure 5 Municipalities in the urban area of Santiago, Chile (black border), and Traffic Analysis Zones under study (grey border), used to assign home and work locations to commuters

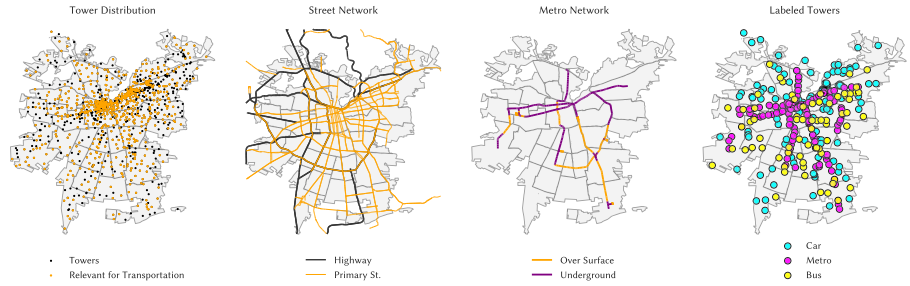

Figure 6 Maps of urban features analyzed in the algorithmic pipeline: the network of cell phone towers in the city, both under consideration and discarded (*e.g.*, indoor towers); the street network, particularly highways and primary streets, from OpenStreetMap (OSM); the metro network, both, underground and over surface, also from OSM; and the set of labeled towers with respect to mode of transportation, according to their proximity to urban features and GTFS

and also improves classification. For instance, due to network congestion it is possible that a device may connect to different towers through several days, even when the trajectory followed by the device is similar. Conversely, the aggregation of towers into zones show less variability. Moreover, since zones respect administrative boundaries, there is a clear mapping between zones and municipalities.

3.3 Urban context: towers, OpenStreetMap and GTFS

The mobile operator has 1374 mobile phone towers in the city. Of them, 787 are relevant for this study, as we discarded towers that were installed in pedestrian streets and indoor contexts (*e.g.*, hospitals, malls, offices, *etc.*). Figure 6 shows the spatial distributions of towers, the street and metro networks of the city, and the association of towers to modes of transportation. The urban networks were obtained from an August 2016

dump of OpenStreetMap (OSM) data. The association to modes of transportation used two sources of information. On the one hand, some towers have meta-data that allow to associate them to urban networks, *e.g.*, underground metro towers have a "Metro" prefix in their names, and some highway towers have "Autopista" (highway) prefix in their names. On the other hand, we associated towers according to their proximity to the urban network obtained from OSM, considering a threshold ratio of 250 meters. For trunk buses with routes in non-bus corridors, we used the GTFS dataset, which contains 376 bus routes. We considered the same threshold of distance to these routes when associating towers.

3.4 Trip inference, home and work location, and OD matrix

The XDR dataset contains all clients with data plans of the mobile operator. Both types of users, contract and pre-paid, are considered. After executing our pipeline, there were 662,665 commuters with a valid modal cluster of pedestrian flag.

Figure 7 displays the spatial distribution of inferred home/work locations in the city. To validate our home location estimation with the CASEN distribution, we estimated the Spearman rank-correlation coefficient ρ, where a value of $\rho = 0$ indicates no correlation; $\rho = 1$ indicates perfect positive correlation of the ranks in the data; and $\rho = -1$ indicates perfect inverse correlation. The result was a coefficient of $\rho = 0.71$ ($p < 0.001$). While the correlation is high, it shows that there is non-negligible deviation. This deviation could be alleviated by weighting users based on their residential distribution. Since the scope of this paper is to provide a method to estimate the modal partition of the city, we leave this weighting for future work.

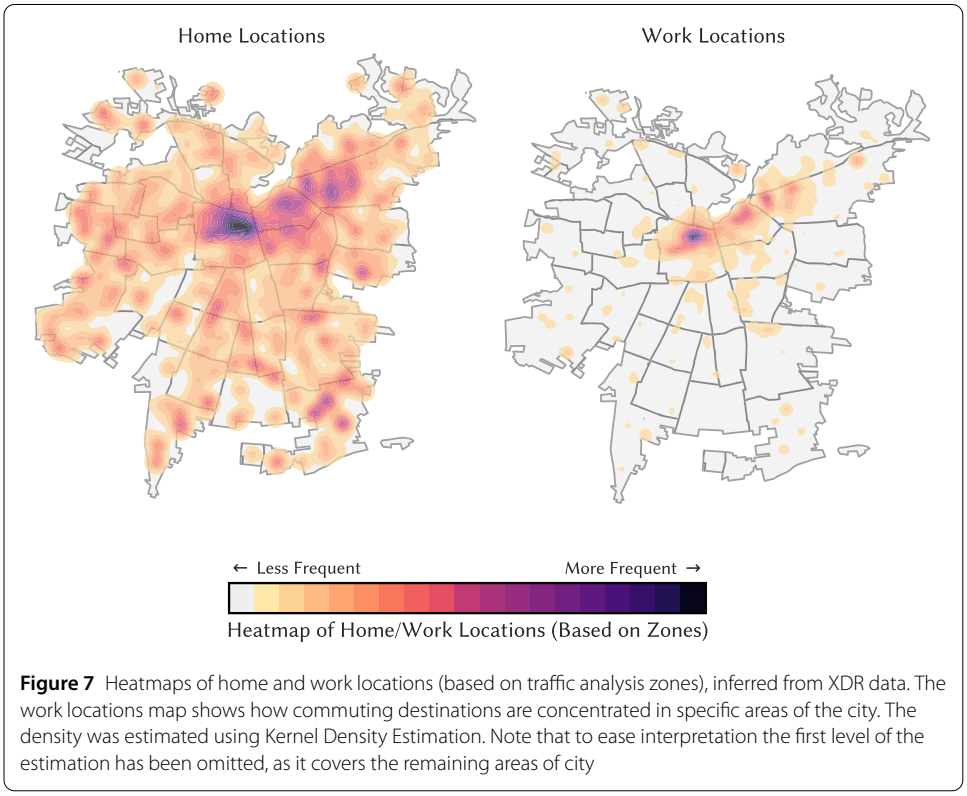

Figure 7 Heatmaps of home and work locations (based on traffic analysis zones), inferred from XDR data. The work locations map shows how commuting destinations are concentrated in specific areas of the city. The density was estimated using Kernel Density Estimation. Note that to ease interpretation the first level of the estimation has been omitted, as it covers the remaining areas of city

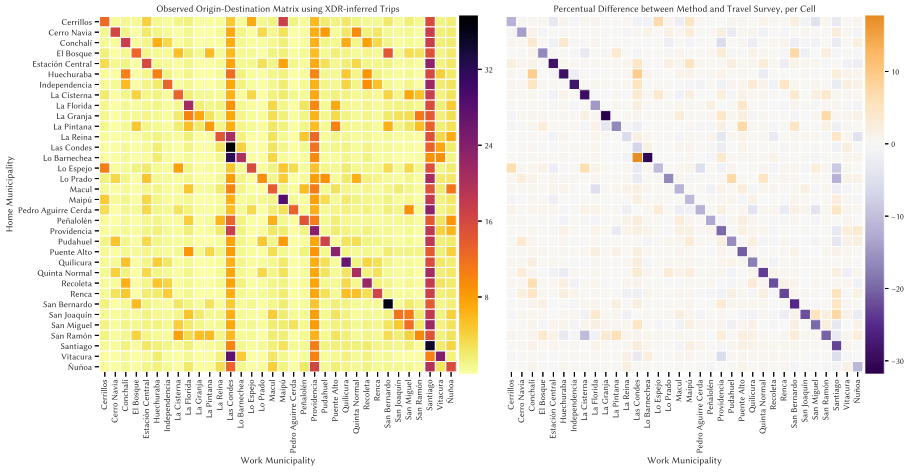

Figure 8 Origin-Destination (OD) matrix from XDR-inferred commuting trips (left). Rows are home municipalities, columns are work municipalities. The matrix on the left is row-normalized and then multiplied by 100 to ease interpretability. The matrix on the right shows the difference with an OD matrix built from the Santiago Travel Survey held in 2012. Each cell in this second matrix is colored with a divergent color palette: grey cells show similar values, orange cells depict OD pairs where XDR infers more trips, and purple cells depict OD pairs where the Travel Survey reports more trips

Regarding work locations, which seem to concentrate on few municipalities, there are several factors that explain this result, including the segregation of the city. These municipalities are characterized by the presence of civic districts, business districts, educational institutions, health institutions, parks and recreation areas, and shopping malls [38]; this explains their prominence on commuting patterns. Destination analyses from other data sources reveal similar results [40, 41]. In fact, when we aggregate commuters into an OD matrix, and compare the results with the matrix from the travel survey, we observe that the differences concentrate on intra-municipality trips. Figure 8 shows this analysis: the matrix on the left is the XDR OD matrix, row-normalized and multiplied by 100. One can see that intra-municipality trips are common, and that three municipalities concentrate a majority of the rest of commuting trips: Santiago Downtown, Providencia, and Las Condes. The matrix on the right depicts the difference between the XDR OD matrix and the one built from the travel survey. Purple cells indicate OD pairs where we observe less trips than in the survey; orange cells, where we observe more trips; and grey cells show a similar proportion of trips. To evaluate how similar are our results to the travel survey, we estimated the Spearman correlation with respect to the cells in which the travel survey reports trips: 904 of 1156 matrix cells. As result, there is a coefficient $\rho = 0.77$ ($p < 0.001$), which indicates that the model results are highly similar to what the survey indicates.

To test whether the distribution of trips is the same, we estimated the Mann–Whitney U statistic, a non-parametric test used to determine whether two independent samples were selected from populations having the same distribution. For the entire set of trips, $U = 325{,}489$ ($p < 0.001$), which means that the distributions are statistically different.

Importantly, it is possible to explain the two most important differences between the two matrices. On the one hand, the diagonal is clearly underrepresented in XDR, mainly

due to the tower distribution. As seen on Fig. 6, the spatial distribution is not uniform, and many municipalities have a low density tower network. The network is denser in places where more people dwell during the day: Santiago Downtown, Providencia, and Las Condes.

4 Modal partition results

In this section we report the results of the later stages of the pipeline. First, we describe the number of labeled users according to their probabilities of using a mode of transportation, $P(u|m)$. In total, we associated 67,559 users to metro, 79,770 users to cars, and 55,545 to buses. As threshold values we used a distinct one for each mode: the 0.9 quantile of the estimated associations according to the *waypoint matrix*. For each user, we updated their prior probability for each mode with the distributions from the CASEN survey. We did so using the Bayes Theorem. As threshold for identification of pedestrian commuting, we considered a maximum distance of one kilometer.

After applying the TS-NMF method, we obtained the A (user-component) and B (tower-component) matrices. Figure 9 displays the spatial distribution of towers according to their association to each component or mode of transportation. One can see that, even though the set of labeled towers was not extensive (*cf.*, Fig. 6), the model propagated the influence of each mode to other towers, by considering the user labels in L. For instance, in case of the metro component, the metro network is clearly depicted, as well as near towers, and towers near bus routes that are used as feeders for metro services.

The matrix A contains the association of users and the mode of transportation components. However, these associations lie in a continuous space, and thus, we needed to discretize them using k-means. We used $k = 4$, with the aim of finding a modal cluster for each mode of transportation, plus one that represents metro and bus. We assumed that *metro and car*, and *car and bus* were not needed, because these are not common combinations of mode of transportation in the city. Figure 10 shows the centroids of each modal cluster.

Figure 11 displays the spatial distribution of home locations, per mode of transportation, based on the modal clusters. Since the home location was performed at the zone level,

Figure 9 Spatial distribution of the column-normalized tower-component matrix B. Each map visualizes a column of the matrix, *i.e.*, the tower associations to each mode of transportation. Each tower is depicted as a purple circle with a size that is proportional to its corresponding association

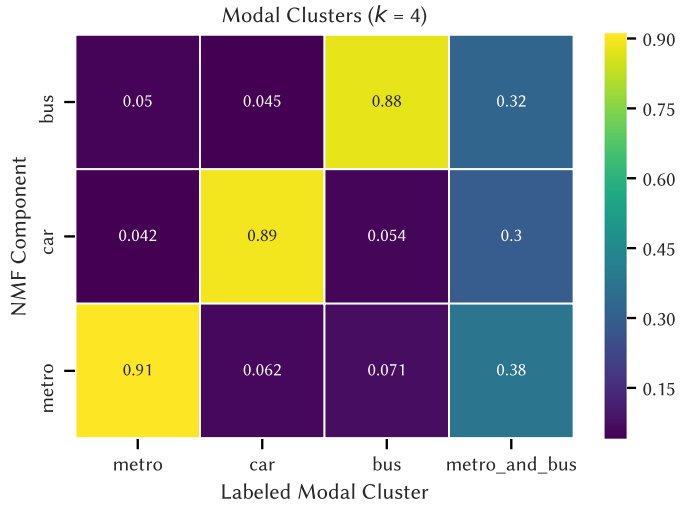

Figure 10 Depiction of the *modal clusters* obtained after applying k-means to the user-component matrix. In this visualization, each column contains the centroid coordinates of its corresponding modal cluster

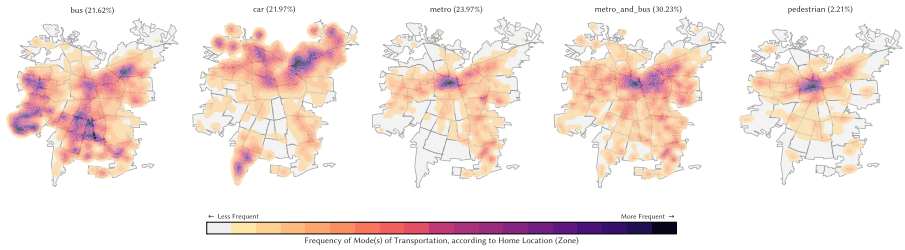

Figure 11 Heatmaps of home location (based on traffic analysis zones) with respect to mode of transportation, estimated with Kernel Density Estimation. Note that to ease interpretation the first level of the estimation has been omitted

one can see subtle differences within municipalities. Figure 12 shows the modal partition per municipality, sorted by car usage. One can see that the shape of the car distribution is similar to the one from CASEN (*cf.*, Fig. 3), however, the order is not the same. The difference in the distribution with respect to home municipalities is shown on Fig. 13.

Figure 14 shows the differences between the inferred OD matrices and the corresponding ones from the Travel Survey. One can see the XDR OD matrix for each modal cluster (top row), the differences with the Travel Survey OD matrix (middle row), and the distribution of the differences using Kernel Density Estimation (bottom row). One can see that the highest correlation with the survey is the one for metro trips, which is expected, as underground stations have their own towers. It is interesting that pedestrian trips have a high correlation too, even though a non-negligible amount of pedestrian trips is not inferred in our data (2.37% *versus* 10.19% prior from CASEN). There are three observable patterns in the difference matrices. First, bus, car, and pedestrian matrices exhibit less trips in the diagonal than in the travel survey. This was discussed earlier, when we compared the global trip matrix. Second, for metro trips, we observe more trips than expected in the Santiago Downtown, Providencia, and Las Condes municipalities. This pattern also repeats from the previous analysis. Third, the metro and bus combination shows the op-

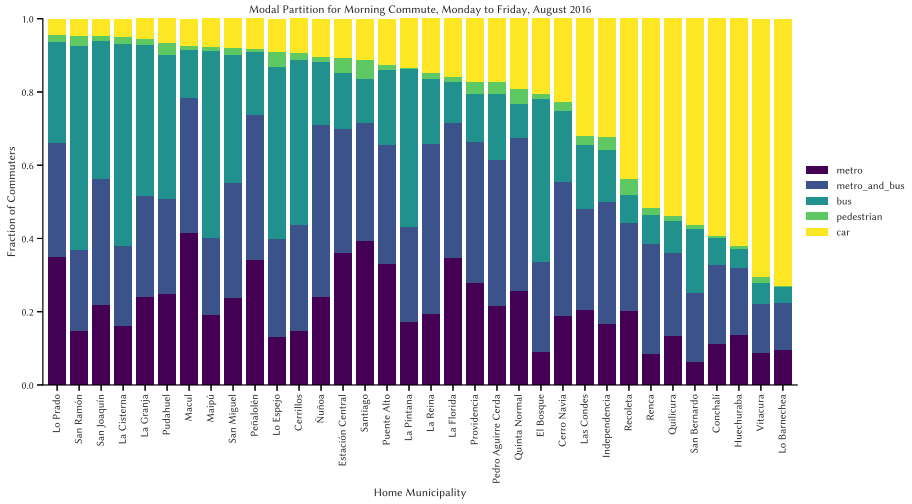

Figure 12 Relative distribution of XDR-inferred mode of transportation usage (*modal partition*) per home municipality in Santiago. The municipalities are sorted with respect to car usage

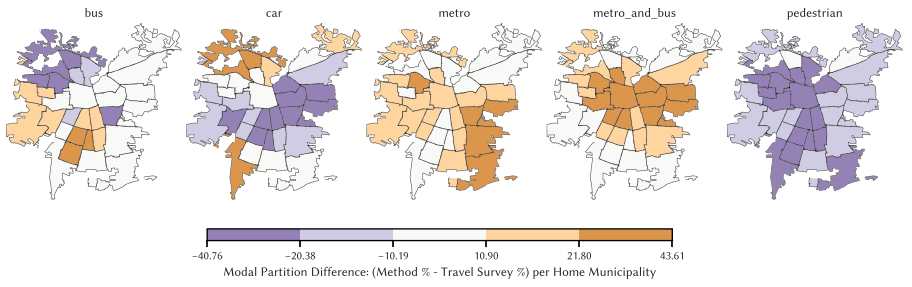

Figure 13 Spatial distribution of the differences of XDR-inferred mode of transportation usage with the modal partition from the Travel Survey. A divergent color palette was used to fill each municipality: grey municipalities exhibit similar values to those from the Travel Survey; orange municipalities have more XDR-inferred trips than in the survey; and purple municipalities have less XDR-inferred trips than in the survey

posite behavior: there are less trips than expected in the three municipalities that attract more trips, and more trips than expected in the diagonal. Arguably, both aspects can be explained. On the one hand, the intra-municipality trips may be shorter, and thus, commuters may have a varied set of public transportation routes to choose from, including bus or metro. On the other hand, the trips to the most frequent destination municipalities may have shifted from intermodal trips to single-modal ones, as Transantiago is designed for multiple boardings, but people may find more value in a single, longer trip, where seat availability is less uncertain [42].

In the bottom row of Fig. 14, we explored the behavior of the model by analyzing the differences as residuals of the pipeline. One can see a distribution with high kurtosis centered in zero; such fat-tailed distributions imply that there are many similar cells in both matrices, but also some with extreme differences.

Table 1 summarizes the results, in terms of modal partition, Spearman rank-correlations, Mann–Whitney *U* statistical tests, and means and standard deviations of residual analysis. As result, the XDR modal partition, even though similar for all modes of transportation

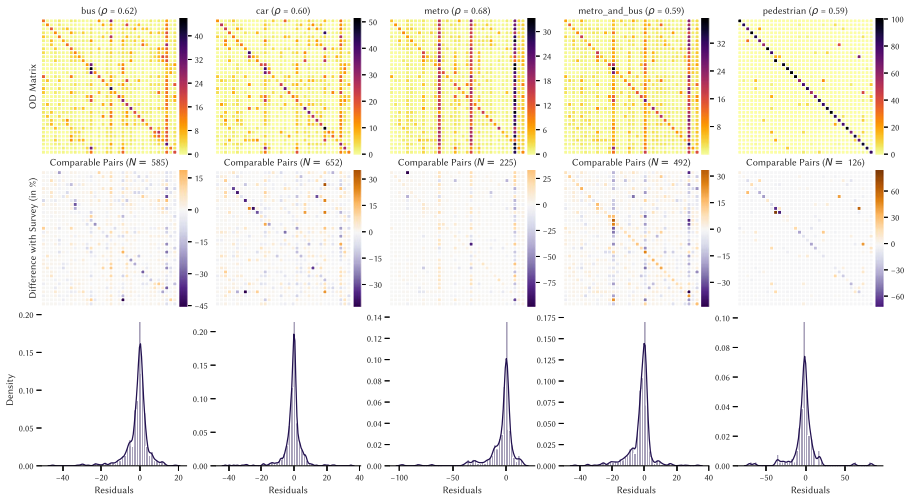

Figure 14 Top row: Origin-Destination (OD) from XDR-inferred trips for each modal cluster. Middle row: difference of XDR-inferred OD matrices with those from the Travel Survey, using a divergent color scale to understand where the matrices differ. Bottom row: distribution of the differences in the middle row, estimated with Kernel Density Estimation

Table 1 Summary results from our proposed pipeline. Each row is a modal cluster or choice of one or more modes of transportation for commuting trips. The Travel Survey column shows the distribution of these modes according to collected data in 2012. The XDR column shows the inferred distribution used our methods. Since the Travel Survey does not contain trips in all possible Origin-Destination (OD) pairs, the third column shows the number of comparable pairs for evaluation. We compared the Spearman rank-correlation coefficient ρ for each modal cluster to evaluate coherence between inferred and known commuting flows, and performed the Mann–Whitney U test to measure whether both distributions were the same. Results indicate a high level of similarity but statistically different distributions. Legend: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$

Mode of Transportation	Travel Survey	XDR	Comparable OD-Pairs	OD ρ	OD U	OD Res. μ	OD Res. σ
Bus	23.69%	21.67%	585	0.63***	171,055	0	4.42
Metro	8.43%	23.90%	225	0.68***	23,692	0.69	7.12
Metro & Bus	14.38%	30.07%	492	0.59***	105,465***	0	5.48
Car	29.60%	22.00%	652	0.60***	198,109*	0	4.82
Pedestrian	23.40%	2.37%	126	0.59***	6408**	0	5.75

(correlations range from 0.59 to 0.68), exhibits different distributions to what is expected from the travel survey, except for unimodal public transportation, where the U test was not significant.

5 Discussion

We presented a method to infer mode of transportation usage in a city, having mobile phone network data as input. We tested it in a big city with an intermodal transportation system, and found that results are coherent with what would be expected from known commuting flows in the city. Care has to be taken when analyzing the differences in the inferred distributions with travel surveys, as surveys have their own drawbacks. As such, we performed a descriptive case study where we provided plausible explanations of these differences.

In terms of implementation for planning purposes, transportation planners work with models that, in addition to trips, work with features at originating areas, such as home locations. In this aspect, we are aligned with their way of work. We have worked closely with the transportation agency from the government to align our methods with their needs, however, there is still a long way ahead, as this pipeline still needs short-term improvements. A first step into extending our pipeline would be the extension for non commuting trips. Other kind of trips are less predictable than commuting; conversely, tower associations do propagate, as the transportation network is the same regardless of trip purpose. Hence, trips can be labeled even if they are not commuting trips, provided that they have enough *within-trip* events that allow to find an association to one of our *modal clusters*. One would ask whether this extension needs as much input data as our case study. We have tested the method with 60K users and two weeks of XDR (as in [19]), and results are comparable. Still, consider that the travel survey expands to almost 3M users with 15K respondents. Moreover, applying the model to an unobserved user is straightforward using a matrix projection. In case of tower network changes, the answer varies. A radical change, for instance, through the installation of new towers within a new metro line, is harder to implement, as it will depend on whether they superimpose with existing towers or not. Incremental update of NMF models is an active research line that could be explored to solve this problem [43].

An important implication of our work is related to explainability and transparency. NMF is considered an explainable model, mainly because of its direct interpretation as learning parts of an object. Parts which, in our case, are the several modes of transportation that comprise commuting trips. In our previous work we showed that plain NMF exhibits this behavior, particularly when compared to Principal Component Analysis [19]. However, our pipeline is more explainable than just NMF, because every step provides an explainable algorithm or method with known units and procedures. Arguably, by checking which OD pairs or municipalities the model differ from the expectations, it is possible to infer which step/parameters need tuning. In our experience, this lack of black boxes has allowed us to communicate the model to transportation experts and to other roles involved in policy making and planning, as well as to define alternative sources of input. For example, if a municipality has a low share of intra-municipality trips, a custom, less expensive survey or a different dataset could be used.

Finally, potential applications of this method include the definition of time-specific OD matrices, which would enable the evaluation of transportation and urban interventions in a city: (i) the estimation of fare evasion in buses, which is high in Santiago [44], by comparing our bus and metro matrices with those derived from smart-card data [40]; (ii) the study of walkability [45], by focusing on the pedestrian trips found; and (iii) the correlation of mode of transportation usage with several urban characteristics, including pollution, safety, crime, among others.

5.1 Future work

We devise two research lines. The first one is the inclusion of bikes into the model. In cities like Santiago the cycling infrastructure is not massive, and thus, labeling towers based on proximity to bikelanes did not produce meaningful results. To be able to include bikes, we could resort to non-traditional data sources like logs from bike applications, but these can only be accessed at each application provider. Furthermore, cyclists may not have regular

routes, and they choice of commuting mode may change day to day due to several factors [46].

The second line of research is the disaggregation of car trips. In its current form, cars include private cars, cabs, and shared cabs, and potentially bikes. It is not clear whether including this information in the prior probabilities is enough, and the raise of ride-hailing applications, such as Uber and Cabify, blur the line in this aspect.

5.2 Scope and limitations

Even though we proposed our model towards transportation planning, critics may rightly say that our results do not conform to those of the travel survey, or that it depends on prior probabilities to be known.

With respect to the comparison with the Travel Survey, we expected that there would be important differences. The tower distribution and the changes in the city since the survey was collected, are factors that the model doest not consider. Additionally, our model is based on assumptions at every level; a relevant one for this discussion is that cars are more prominent in secondary streets than feeder buses. This is arguably true, particularly in richer areas, and where it is not, the prior probabilities solve the problem. Thus, the dependence on prior probabilities is not actually a limitation, but a feature of the model that allows to include such assumptions. Perhaps a limitation would be the actual estimation of such probabilities. In this work we built them from a survey with municipal representativity, that is by itself costly. However, since we just need the mode of transportation distribution, the prior probabilities could be manually specified by transportation experts, or through non-traditional datasets, such as smart-card data [40]. Other approaches would include the use of scaling parameters to solve the issue [47].

A potential source of bias is the market share of the mobile operator. Since the data is anonymized, it is not possible to weight users according to their socio-economic status. To solve this, we estimated the modal partition at the municipality level, and performed our analysis with row-normalized matrices. As the scope of this paper is to provide a method to infer mode of transportation, rather than generate a balanced dataset, we leave this for future work when implementing our proposal within applied contexts.

6 Related work

Mobile phone network data (XDR and other types) has enabled a flurry of studies about the laws behind behind human mobility [48, 49], as well as behavioral and societal analyses [11]. The interest in XDR is not only theoretical, it also provides a cost-effective way of understanding behavior in developing and emerging countries [50], which may not have institutional or economical means to gather data.

In our context, mobile phone network data has been used to infer trips and to aggregate them into OD matrices [18, 20, 22, 51, 52]. Some of these approaches go as far as expanding the sample to be representative of the population, for instance, by incorporating other sources of data, such as traffic counts [51]. However, XDR has not been used to effectively include mode(s) of transportation into OD matrices. On the one hand, current approaches to infer mode(s) of transportation make naive assumptions; for instance, it is assumed that trips are unimodal (*i.e.*, one mode of transportation per trip), that there are only two types of motorized transportation (car and public transportation, usually bus), and that average vehicle speed follows a bimodal distribution [53, 54]. The bimodality assumption is feasible [55], however, it has two important problems. On the one hand, it is

not applicable in congested cities, in cities with more than two main modes of transportation (*e.g.*, availability of metro), or where intermodality in trips is common. The variability of travel times is a current research line for transportation experts [56], showcasing how such assumptions based on time (and, thus, on speed) may not be desirable for complex, growing cities like Santiago. On the other hand, XDR data is comprised by several types of billing records, including several granularity levels. For instance, some XDR datasets include triangulated device positions instead of towers; others may have temporal delays in event timestamps [23]. In our case, we use an arguably low granularity XDR stream: an average of 15 minutes between events, and tower-based positions. This granularity hinders the estimation of variables that are used in previous works: travel time [53] and speed [54].

A different way of approaching the problem of mode inference is by analyzing trip routes. Using GPS data, the inference of mode(s) of transportation is a solved problem, as it provides acceleration patterns and followed routes [57, 58]. GPS has a spatio-temporal granularity that is orders of magnitude finer than XDR, as it is collected many times per second, allowing the usage of methods like *map-matching* between points and transportation infrastructure [59]. Although available on almost any commodity smartphone, GPS drains battery, it is not always activated, and requires specific applications to be installed to gather data. Another approach is the usage of sensor data from mobile phones, which has similar limitations to that from GPS. For instance, accelerometer data allows to identify mode of transportation, as different modes exhibit different patterns of acceleration [60]. XDR, in contrast, is *passive data*, in the sense that it is always generated regardless of user actions; it does not reveal specific locations, but areas of tower coverage; and it is cost-effective, as it is already generated and stored.

Hence, there is a limitation in the current state of the art regarding inference of mode of transportation using XDR. To improve this state, we focused on commuting, which represents a major portion of the trips within a city, and it is a recurrent trip that allows to aggregate tower connectivity. The core method in our proposal is the usage of a dimensionality reduction technique called Non-Negative Matrix Factorization (NMF) [32], that is equivalent to performing spectral clustering [30] of the two different types of entities in our data: devices (and, by extension, their users) and cell phone towers. NMF is interpretable, as it learns how to separate objects into a sum of its parts [26]. It has been applied in computational biology [61], urban analysis [62], and, with XDR, on trip purpose inference [63]. As such, our hypothesis was that NMF clusters aggregated trajectories, allowing to interpret them as a sum of the chosen modes of transportation. In our previous work, we found that NMF finds clusters that are spatially separable, in contrast to those found with other techniques such as Principal Component Analysis [19]. Some of these clusters were related to transportation infrastructure; here we proposed to guide the learning process toward transportation clusters only through Topic-Supervised NMF [16].

Even though we have centered the discussion around mobile phone network data, it is possible to infer transportation and urban patterns from other kinds of datasets: smartcard data [40, 64, 65]; Twitter, which has been shown to be a good predictor of commuter flows [66] and mobility patterns [67]; and Flickr, that has been used to fit mobility models [68]. Indeed, any data source that allows to count the number of people that goes from one place to another can be used to fit gravity or radiation models (see [69] for a comparison). The limitation, in our context, is that such models do not consider within-trip waypoints,

and thus, limit the inference of mode of transportation to variables that may not be reliable, such as speed or travel time.

7 Conclusions

In this paper we presented an interdisciplinary approach to infer the distribution of mode of transportation usage for commuting. The approach follows the conventions and parameters of the area of application, Transportation [13], and uses tools and methods from Data Science [12] applied to a non-traditional data source: billing records from mobile phone networks. By performing a case study in a big city, Santiago, we found that the proposed method delivers coherent results, as all modes of transportation under study exhibit similar rank-correlations with the travel survey (from 0.59 to 0.68). Furthermore, our algorithmic pipeline is explainable, in the sense of being able to associate differences in the results with those from a travel survey with specific steps and parameters. Given that the current source of this kind of insight for transportation experts are surveys, that may be outdated, we believe that our work contributes to both disciplines, Data Science and Transportation. Finally, by considering the proposed methods and its results, domain experts will be able to augment their work in a cost-effective way by performing a finer analysis of how people lives and moves in their cities.

Acknowledgements

The analysis was performed using Jupyter Notebooks [70], jointly with the *scikit-learn* [71], *pandas* [72], *geopandas*, *statsmodels* [73], *seaborn*, *ws-nfm* [16], and *imposm.parser* libraries. The maps on this paper include data from ©OpenStreetMap contributors. We thank Telefónica R&D in Santiago for facilitating the data for this study, in particular Pablo García Briosso. We also thank Ciro Cattuto (ISI), Leo Ferres (UDD), Ricardo Hurtubia (PUC), Viviana Muñoz (SECTRA), and Marcelo Tapia (UDD) for valuable discussion. The authors E. Graells-Garrido and D. Caro acknowledge financial support from the Chilean government initiative CORFO 13CEE2-21592 (2013-21592-1-INNOVA_PRODUCCION2013-21592-1). The author D. Parra has been funded by Conicyt, Fondecyt grant 11150783, as well as Fondef grant id16i10222 and the BRT+ Centre of Excellence funded by VREF.

Abbreviations

CASEN, *CAracterización Socio-Económica* (Socio-Economic Characterization); GPS, Global Positioning System; GTFS, General Transit Feed Specification; OD, Origin-Destination; OSM, OpenStreetMap; NMF, Non-negative Matrix Factorization; TS-NMF, Topic-Supervised NMF; XDR, Data Detail Records.

Availability of data and materials

The Telefónica Movistar mobile phone records have been obtained directly from the mobile phone operator through an agreement between the Data Science Institute and Telefónica R&D. This mobile phone operator retains ownership of these data and imposes standard provisions to their sharing and access which guarantee privacy. Anonymized datasets are available from Telefónica R&D Chile (http://www.tidchile.cl) for researchers who meet the criteria for access to confidential data. The other datasets used in this study are publicly available at the following addresses: CASEN 2015 (http://observatorio.ministeriodesarrollosocial.gob.cl/casen-multidimensional/casen/basedatos.php), Travel Survey 2012 (http://datos.gob.cl/dataset/31616), GTFS (http://datos.gob.cl/dataset/33245), and OpenStreetMap data dump (http://download.geofabrik.de/south-america/chile.html).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

EG and DC defined the conceptual pipeline and performed data analysis. EG was mainly responsible for NMF model implementation, evaluation, visualization, and tower labeling; DC was mainly responsible for the implementation of home/work detection, and data cleaning and processing. EG and DP participated in manuscript preparation. All authors read and approved the final manuscript.

Author details

[1]Data Science Institute, Faculty of Engineering, Universidad del Desarrollo, Santiago, Chile. [2]Telefonica R&D, Santiago, Chile. [3]School of Engineering, Pontificia Universidad Catolica, Santiago, Chile.

Publisher's Note

Received: 16 August 2018 Accepted: 22 November 2018 Published online: 04 December 2018

References

1. Montgomery C (2013) Happy city: transforming our lives through urban design. Macmillan Co., New York
2. Lyons G, Chatterjee K (2008) A human perspective on the daily commute: costs, benefits and trade-offs. Transp Rev 28(2):181–198
3. Rüger H, Pfaff S, Weishaar H, Wiernik BM (2017) Does perceived stress mediate the relationship between commuting and health-related quality of life?. Transp Res Part F Traffic Psychol Behav 50:100–108
4. Kahneman D, Krueger AB, Schkade DA, Schwarz N, Stone AA (2004) A survey method for characterizing daily life experience: the day reconstruction method. Science 306(5702):1776–1780
5. González F, Melo-Riquelme C, de Grange L (2016) A combined destination and route choice model for a bicycle sharing system. Transportation 43(3):407–423. https://doi.org/10.1007/s11116-015-9581-6
6. Stewart DW, Shamdasani PN (2014) Focus groups: theory and practice, vol 20. Sage, Thousand Oaks
7. Cascetta E (1984) Estimation of trip matrices from traffic counts and survey data: a generalized least squares estimator. Transp Res, Part B, Methodol 18(4):289–299
8. Kuwahara M, Sullivan EC (1987) Estimating origin-destination matrices from roadside survey data. Transp Res, Part B, Methodol 21(3):233–248
9. Groves RM (2006) Nonresponse rates and nonresponse bias in household surveys. Public Opin Q 70(5):646–675
10. Calabrese F, Ferrari L, Blondel VD (2015) Urban sensing using mobile phone network data: a survey of research. ACM Comput Surv 47(2):25
11. Blondel VD, Decuyper A, Krings G (2015) A survey of results on mobile phone datasets analysis. EPJ Data Sci 4(1):1
12. Cao L (2017) Data science: a comprehensive overview. ACM Comput Surv 50(3):43
13. Hall R (2012) Handbook of transportation science, vol 23. Springer, New York
14. Yates RB, Neto BR (2011) Modern Information Retrieval: the concepts and technology behind search. Addison-Wesley Professional
15. Cichocki A, Phan A-H (2009) Fast local algorithms for large scale nonnegative matrix and tensor factorizations. IEICE Trans Fundam Electron Commun Comput Sci 92(3):708–721
16. MacMillan K, Wilson JD (2017) Topic supervised non-negative matrix factorization. arXiv preprint. arXiv:1706.05084
17. Urner R, David SB, Shamir O (2012) Learning from weak teachers. In: Artificial intelligence and statistics, pp 1252–1260
18. Graells-Garrido E, Saez-Trumper D (2016) A day of your days: estimating individual daily journeys using mobile data to understand urban flow. In: Proceedings of the second international conference on IoT in urban space. ACM, New York, pp 1–7
19. Graells-Garrido E, Caro D, Parra D (2018) Toward finding latent cities with non-negative matrix factorization. In: Said A, Komatsu T (eds) Workshop on user interfaces for spatial-temporal data analysis. http://ceur-ws.org/Vol-2068/uistda4.pdf
20. Calabrese F, Di Lorenzo G, Liu L, Ratti C (2011) Estimating origin-destination flows using mobile phone location data. IEEE Pervasive Comput 10(4):0036
21. Visvalingam M, Whyatt JD (1993) Line generalisation by repeated elimination of points. Cartogr J 30(1):46–51
22. Alexander L, Jiang S, Murga M, González MC (2015) Origin–destination trips by purpose and time of day inferred from mobile phone data. Transp Res, Part C, Emerg Technol
23. Graells-Garrido E, Peredo O, García J (2016) Sensing urban patterns with antenna mappings: the case of Santiago, Chile. Sensors 16(7):1098
24. Thai J, Laurent-Brouty N, Bayen AM (2016) Negative externalities of gps-enabled routing applications: a game theoretical approach. In: Intelligent transportation systems (ITSC), 2016 IEEE 19th international conference on. IEEE Press, New York, pp 595–601
25. Lee DD, Seung HS (2001) Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems, pp 556–562
26. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401(6755):788–791
27. Cullum J, Willoughby RA, Lake M (1983) A lanczos algorithm for computing singular values and vectors of large matrices. SIAM J Sci Stat Comput 4(2):197–215
28. Pearson K (1901) LIII. On lines and planes of closest fit to systems of points in space. London, Edinburgh, Dublin Philos Mag J Sci 2(11):559–572
29. Xu W, Liu X, Gong Y (2003) Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, pp 267–273
30. Ding C, He X, Simon HD (2005) On the equivalence of nonnegative matrix factorization and spectral clustering. In: Proceedings of the 2005 SIAM international conference on data mining. SIAM, Philadelphia, pp 606–610
31. Gaussier E, Goutte C (2005) Relation between plsa and nmf and implications. In: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, pp 601–602
32. Kuang D, Ding C, Park H (2012) Symmetric nonnegative matrix factorization for graph clustering, pp 106–117
33. Kim J, Park H (2008) Toward faster nonnegative matrix factorization: a new algorithm and comparisons. In: Data mining, 2008. ICDM'08. Eighth IEEE international conference on. IEEE Comput. Soc., Los Alamitos, pp 353–362
34. Sculley D (2010) Web-scale k-means clustering. In: Proceedings of the 19th international conference on World Wide Web. ACM, New York, pp 1177–1178
35. Weinstein Agrawal A, Schlossberg M, Irvin K (2008) How far, by which route and why? A spatial analysis of pedestrian preference. J Urban Des 13(1):81–98
36. O'Sullivan S, Morrall J (1996) Walking distances to and from light-rail transit stations. Transp Res Rec 1538:19–26
37. Graells-Garrido E, Ferres L, Caro D, Bravo L (2017) The effect of Pokémon Go on the pulse of the city: a natural experiment. EPJ Data Sci 6(1):23
38. Beiró MG, Bravo L, Caro D, Cattuto C, Ferres L, Graells-Garrido E (2018) Shopping mall attraction and social mixing at a city scale. EPJ Data Sci 7(1):28
39. Muñoz JC, Gschwender A (2008) Transantiago: a tale of two cities. Res Transp Econ 22(1):45–53

40. Munizaga MA, Palma C (2012) Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from Santiago, Chile. Transp Res, Part C, Emerg Technol 24:9–18
41. Kickhofer B, Hosse D, Turnera K, Tirachini A (2016) Creating an open matsim scenario from open data: the case of Santiago de Chile. Technical report, VSP Working Paper 16-02
42. Arentze TA, Molin EJ (2013) Travelers' preferences in multimodal networks: design and results of a comprehensive series of choice experiments. Transp Res, Part A, Policy Pract 58:15–28
43. Chen X, Candan KS (2014) Gi-nmf: group incremental non-negative matrix factorization on data streams. In: Proceedings of the 23rd ACM international conference on conference on information and knowledge management. ACM, New York, pp 1119–1128
44. Guarda P, Galilea P, Paget-Seekins L, de Dios Ortúzar J (2016) What is behind fare evasion in urban bus systems? An econometric approach. Transp Res, Part A, Policy Pract 84:55–71
45. Quercia D, Aiello LM, Schifanella R, Davies A (2015) The digital life of walkable streets. In: Proceedings of the 24th international conference on World Wide Web, pp 875–884. International World Wide Web Conferences Steering Committee
46. Heinen E, Maat K, Van Wee B (2011) Day-to-day choice to commute or not by bicycle. Transp Res Rec 2230:9–18
47. Yang Y, Herrera C, Eagle N, González MC (2014) Limits of predictability in commuting flows in the absence of data for calibration. Sci Rep 4:5662
48. Gonzalez MC, Hidalgo CA, Barabasi A-L (2008) Understanding individual human mobility patterns. Nature 453(7196):779–782
49. Candia J, González MC, Wang P, Schoenharl T, Madey G, Barabási A-L (2008) Uncovering individual and collective human dynamics from mobile phone records. J Phys A, Math Theor 41(22):224015
50. Hilbert M (2016) Big data for development: a review of promises and challenges. Dev Policy Rev 34(1):135–174
51. Iqbal MS, Choudhury CF, Wang P, González MC (2014) Development of origin–destination matrices using mobile phone call data. Transp Res, Part C, Emerg Technol 40:63–74
52. Frias-Martinez V, Soguero C, Frias-Martinez E (2012) Estimation of urban commuting patterns using cellphone network data. In: Proceedings of the ACM SIGKDD international workshop on urban computing. ACM, New York, pp 9–16
53. Wang H, Calabrese F, Di Lorenzo G, Ratti C (2010) Transportation mode inference from anonymized and aggregated mobile phone call detail records. In: Intelligent transportation systems (ITSC), 2010 13th international IEEE conference on. IEEE Press, New York, pp 318–323
54. Qu Y, Gong H, Wang P (2015) Transportation mode split with mobile phone data. In: Intelligent transportation systems (ITSC), 2015 IEEE 18th international conference on. IEEE Press, New York, pp 285–289
55. Glaeser EL, Kahn ME, Rappaport J (2008) Why do the poor live in cities? The role of public transportation. J Urban Econ 63(1):1–24
56. Durán-Hormazábal E, Tirachini A (2016) Estimation of travel time variability for cars, buses, metro and door-to-door public transport trips in Santiago, Chile. Res Transp Econ 59:26–39
57. Zheng Y, Zhang L, Xie X, Ma W-Y (2009) Mining interesting locations and travel sequences from gps trajectories. In: Proceedings of the 18th international conference on World Wide Web. ACM, New York, pp 791–800
58. Feng T, Timmermans HJ (2013) Transportation mode recognition using GPS and accelerometer data. Transp Res, Part C, Emerg Technol 37:118–130
59. Quddus M, Washington S (2015) Shortest path and vehicle trajectory aided map-matching for low frequency gps data. Transp Res, Part C, Emerg Technol 55:328–339
60. Jahangiri A, Rakha HA (2015) Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. IEEE Trans Intell Transp Syst 16(5):2406–2417
61. Devarajan K (2008) Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. PLoS Comput Biol 4(7):1000029
62. Wakamiya S, Lee R, Kawai Y, Sumiya K (2015) Twitter-based urban area characterization by non-negative matrix factorization. In: Proceedings of the 2015 international conference on big data applications and services. ACM, New York, pp 128–135
63. Peng C, Jin X, Wong K-C, Shi M, Liò P (2012) Collective human mobility pattern from taxi trips in urban area. PLoS ONE 7(4):34487
64. Caminha C, Furtado V, Pinheiro V, Silva C (2016) Micro-interventions in urban transportation from pattern discovery on the flow of passengers and on the bus network. In: Smart cities conference (ISC2), 2016 IEEE international. IEEE Press, New York, pp 1–6
65. Alsger A, Assemi B, Mesbah M, Ferreira L (2016) Validating and improving public transport origin–destination estimation algorithm using smart card fare data. Transp Res, Part C, Emerg Technol 68:490–506
66. McNeill G, Bright J, Hale SA (2017) Estimating local commuting patterns from geolocated Twitter data. EPJ Data Sci 6(1):24
67. Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C (2014) Geo-located Twitter as proxy for global mobility patterns. Cartogr Geogr Inf Sci 41(3):260–271
68. Beiró MG, Panisson A, Tizzoni M, Cattuto C (2016) Predicting human mobility through the assimilation of social media traces into mobility models. EPJ Data Sci 5(1):30
69. Masucci AP, Serras J, Johansson A, Batty M (2013) Gravity versus radiation models: on the importance of scale and heterogeneity in commuting flows. Phys Rev E 88(2):022812
70. Pérez F, Granger BE (2007) IPython: a system for interactive scientific computing. Comput Sci Eng 9(3):21–29
71. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
72. McKinney W (2010) Data structures for statistical computing in Python. In: Proceedings of the 9th Python in science conference, vol 445, pp 51–56
73. Seabold S, Perktold J (2010) Statsmodels: econometric and statistical modeling with Python. In: Proceedings of the 9th Python in science conference, vol 57, p 61. SciPy society Austin