A Generalisable Data Fusion Framework to Infer Mode of Transport Using Mobile Phone Data

Eduardo Graells-Garrido†  Daniela Opitz†  Francisco Rowe‡

† Data Science Institute, Faculty of Engineering, Universidad del Desarrollo, Chile
‡ Department of Geography and Planning, University of Liverpool, UK

*Correspondence: Eduardo Graells-Garrido <egraells@udd.cl>

Abstract

Cities often lack up-to-date data analytics to evaluate and implement transport planning interventions to achieve sustainability goals, as traditional data sources are expensive, infrequent, and suffer from data latency. Mobile phone data provide an inexpensive source of geospatial information to capture human mobility at unprecedented geographic and temporal granularity. This paper proposes a method to estimate updated mode of transportation usage in a city, with novel usage of mobile phone application traces to infer previously hard to detect modes, such as bikes and ride-hailing/taxi. By using data fusion and matrix factorisation, we integrate socioeconomic and demographic attributes of the local resident population into the model. We tested the method in a case study of Santiago (Chile), and found that changes from 2012 to 2020 in mode of transportation inferred by the method are coherent with expectations from domain knowledge and the literature, such as ride-hailing trips replacing mass transport.

Introduction

Understanding travel behaviour is key to creating more resilient and sustainable urban transport networks. The global collective contribution of the transport sector to carbon emissions has been estimated to be over 14% of the global estimate (Allen et al., 2012). As cities expand and absorb population growth, global warming (Acar and Dincer, 2020), air pollution (Colvile et al., 2001), traffic congestion (Thomson, 1998), and urban sprawl (Brueckner, 2005), among other issues, exert increasing pressure on decarbonising transport networks and more sustainable modes of travel.

However, cities often lack suitable real- or near real-time data analytics to evaluate and implement transport planning interventions to achieve these goals. Traditional data sources, such as travel surveys, manual count studies, and censuses, have conventionally been used to estimate traffic counts and assess the spatial patterns of human mobility and how people move within cities (Zannat and Choudhury, 2019). Yet, these data sources are expensive, infrequent, and suffer from data latency (Rowe, 2022). Thus, while these data may provide an accurate representation of urban transport networks at a given point in time, such rendering of transport patterns may be unreliable or outdated a few years into the future, and this may rapidly occur in fast-expanding cities or as they may experience shocks due to natural hazards, pandemics or social unrest.

Digital trace data collected through mobile phones can help overcome these deficiencies (Rowe, 2022). Mobile phone data provide an inexpensive source of geospatial information to capture human mobility at unprecedented geographic and temporal granularity (Zannat and Choudhury, 2019). Mobile phone data also offer information to capture travel behaviour in real or near-real time, providing an opportunity to frequently updates as to the ways people move around cities to support transport planning (Antoniou et al., 2011). Despite these major advantages, mobile phone data are limited, unequally and systematically representing certain population groups (Rowe, 2022). They only provide geospatial information on users but not personal attributes, such as age, gender and income which are often associated with different travel behaviours. Additionally, mobile phone data do not provide information on mode of transportation. This is a key piece of information
This paper aims to develop a generalisable Data Fusion framework to update existing area-level mode of transport (defined as mode split) estimates by leveraging the strength of traditional and new forms of data. Data Fusion involves the integration of multiple related datasets into a unified representation that facilitates the extraction of knowledge and patterns from it. Data Fusion in transportation has multiple uses (El Faouzi et al., 2011), mainly due to the flurry of digital information sources that can aid planning, such as traffic counters and other types of sensors. Specifically, we seek to update these estimates from travel survey data by integrating origin-destination mode of transport distributions from more up-to-date mobile phone data from Deep Packet Inspection (DPI) and Extended Detail Records (XDR), and socioeconomic and demographic information from a representative household survey and census data. XDR and DPI data are used to infer a more up-to-date origin-destination mode of transport estimate. XDR provide geospatial locations of users, and DPI data offer information on application usage of related transport modes to infer these estimates based on existing data. Household survey and census data are used to incorporate data on the socio-economics and demographics of the resident population and adjust biases for distributional biases relating to these attributes in mobile phone data.

The paper sits at the intersection of two distinct branches of research: data fusion and mode-of-transport inference from mobile phone data. Our approach builds on existing work inferring modes of transport from mobile phone data (Bachir et al., 2019; Graells-Garrido, Caro and Parra, 2018), which essentially group mobility data and infer the split mode of different cities by using clustering algorithms. We expand these approaches by incorporating data: (1) from mobile phone application usage to infer the use of bike-sharing platforms and taxi services to improve estimates of bicycle and taxi transport modes; and (2) socioeconomic and demographic attributes of the local resident population to integrate information about variation in the usage of different transport modes.

Background

For transport mode detection, most of the work has been done using GPS data (e.g., Bantis and Haworth (2017), Dabiri and Heaslip (2018)), as these offer high spatial and temporal precision. However, as GPS data are collected by smartphone applications, they are restricted to a small sample of the population (Huang et al., 2019). Mobile network data collected by network operators, in contrast, potentially enable the study of the whole population mobility. Despite this advantage, in the context of transport mode detection, we can find few works using mobile network data, mainly based on non-supervised or semi-supervised machine learning algorithms. The main difficulty in developing models to infer the modal split based on supervised algorithms using mobile phone data is the lack of labelled data.

From the few works available, we highlight three from the last decade. First, Wang et al. (2010) use a non-supervised machine learning algorithm to cluster travel times from mobile phone data and identify two modes of transport: road and public transport, in the city of Boston. Second, Graells-Garrido, Caro and Parra (2018) propose a method based on semi-supervised Non-negative Matrix Factorisation (NMF; Kuang et al. (2012)) to infer the transportation modal split of Santiago of Chile, distinguishing the use of rail, car, bus, rail+bus, and pedestrian modes. The essential idea of the NMF is to decompose a high-dimensional matrix W into the factorisation of two matrices, A and B, to identify a set of non-negative components (or latent factors) of an object. In the referenced work, W corresponds to a matrix of the frequencies of start trips of users (rows) at mobile service towers (columns), A encodes k latent features or transport modes (columns) for users (rows), and B encodes k modes of transport (rows) for towers (columns). Third, Bachir et al. (2019) use agglomerative hierarchical clustering to group Voronoi sectors of Paris with similar transport usage distinguishing the modes: rail, road, and multimodal (raillroad); and rescaling OD flows with expansion factors calculated with mobile network data and census.

Arguably, these methods are similar, as NMF is equivalent to performing clustering of the two different types of entities in mobile phone data: devices (and, by extension, their users) and towers (Ding et al., 2005). However, these works have different limitations. The work of Wang et al. (2010) can distinguish just two transport modes.
and is based only on travel times, which can vary at different traffic stages and could be similar under certain conditions, particularly in congested cities. The model of Graells-Garrido, Caro and Parra (2018), in addition to the modes of road and public transport, can infer rail, pedestrian mode, and the intermodal mode rail+bus, though the results of the pedestrian mode are underestimated. On the other hand, the model of Bachir et al. (2019) only considers a bimodal partition (rail and road) and does not consider the intermodality of trips.

However, when more datasets are available, the NMF method does not allow for directly incorporating multiple sources of information and cannot take advantage of the potential relations between multiple datasets. Integrating more than two datasets can be addressed using a feature-based machine learning algorithm, but it requires a single feature-based table, which is sometimes impractical to build. A better approach is the use of data fusion algorithms, which are explicitly designed to address the multiplicity of data and fuse them through inference of a single joint model. Data fusion algorithms have been used to address a variety of problems, mainly in the biomedical research area (Li et al., 2018) and the modelling of urban lifestyles of individuals (González, 2019), but have not been used on the mode split problem yet.

To respond to the need to incorporate external information into urban models, we applied a collective matrix factorisation method based on data fusion (Žitnik and Zupan, 2014) to infer the modal split in Santiago. This model allows for linking several datasets and factorising each matrix in the data, maintaining the coherence of how the different entities modelled in each matrix appear in different datasets.

**Methods and Data**

This work aims to provide a method to estimate an updated mode split of a city, defined as the distribution of the usage of the mode of transportation through the city. It is commonly analysed in two levels: total or city level, and municipal level if the city has several municipalities or districts. Mode split estimations usually come from traditional data sources, mainly travel surveys and population censuses.

Digital traces are non-representative of the population; however, they contain signals of population behaviour at greater resolution than traditional data, both spatial and temporal. As such, we propose a method that takes as input both types of data, traditional and digital sources, and integrates them so that an updated modal split can be estimated. We use a data fusion approach, a suite of techniques to integrate data using a unified representation. We focus on data fusion through matrix factorisation, which enables us to learn a latent representation of the unified data (Žitnik and Zupan, 2014). Latent representations can be used to find patterns in the data (such as clusters), to provide a compressed view of the data (such as dimensional reduction), and to reconstruct original data (such as matrix completion).

For Santiago we used the last official travel survey (2012) to build an initial mode split estimation, with attributes such as Origin-Destination (OD) flows and distributions of travel distance and speed per mode of transportation. We fused this data with more recent official data, including a socio-economic characterisation survey and a national census from 2017, containing attributes such as age, sex income, the number of mobile phone users, the household car ownership distribution, and household income at different geographical resolutions. Although these official data are representative of the population, these do not allow an estimate of an updated mode split. More recent data comes from digital traces: OD flows and application usage from mobile phone data in 2020. All these data are represented in a unified network of matrices that are processed using a matrix factorisation method. Part of the digital data is related to the mode split: OD flows can be analog between datasets, and mobile phone applications can be mapped to modes of transportation. Thus, when reconstructing the modal split, we expect the model to consider the new patterns learned from the data. The result is an updated mode split (Figure S1 shows a schema of the process).

Next, we detail this process. First, we describe the context of our case study and the data used, and then we describe the details of our methodology, including the data fusion approach (Žitnik and Zupan, 2014).

**Context and Data**

Our work studies the Great Metropolitan Area of Santiago, Chile. Santiago is a city with almost 8 million inhabitants and an urban surface of 867.75 square kilometres as of 2017, composed of more than 40 independent
Santiago has an integrated multimodal system. Trips can mix metro, bus, and train through a smart card. The city also has ubiquitous taxi services and access to several ride-hailing applications, and a public bike-sharing system. For this work, we congregate all modes of transportation into four categories: **mass-transport**, which includes all public transportation that uses the city smart card; **private**, which refers to motorised transport; **non-motorised**, which includes pedestrian and cycle trips; and **taxi**, which includes traditional taxi and ride-hailing applications. The last official mode split for Santiago was published in 2015, based on a travel survey from 2012. The survey is outdated considering the different interventions and changes in the city, making it difficult to understand current travel behaviour patterns. We emphasise three of these changes. First, the population composition of the city has significantly changed due to migration waves. For instance, according to the last Chilean census, 61% of the migrant population arrived in the country between 2015 and 2017. Second, new transportation infrastructure was built, including two metro lines and a new train service from 2017 onwards, and new mobile phone applications and other transportation systems appeared. And third, the social upheaval from October 2019 caused multiple riots throughout the city, including burned metro stations (Urquieta Ch., 2019), the closing of commercial districts, and an associated economic crisis that was deepened during the COVID-19 pandemic (Somma et al., 2021).

To estimate an updated mode split, we integrated the following data sources about Santiago: mobile phone data from 2020, official socioeconomic characterisation, census from 2017, the travel survey collected in 2012, and sociodemographic open data sources. Next, we describe these sources.

### Mobile Phone Data

We used two mobile phone datasets provided by the telecommunications operator Telefonica Movistar in Chile: aggregated trajectories from XDR, and aggregated traffic from Deep Packet Inspection (DPI). Both datasets were generated from 749K phones active in the study area on March 5th, 2020 (a Thursday). Telefónica has nearly one-third of the market in the country. In the urban area of Santiago operates 2076 towers distributed in the area of study (see Figure 1.a). We studied municipalities with at least 5 towers in the area under study, resulting in 40 municipalities. We focused on the morning commute time (from 6 AM to 9 AM), as it is the most generalisable period that can be studied during that day.

The primary unit of analysis in the XDR dataset is a billing record (billable events include calls, SMS, and Internet downloads). On average, they are generated in 15-minute intervals for each device (Graells-Garrido et al., 2016). Each record contains the ID of the tower the device was connected to, a timestamp, and an
anonymised device ID. Device IDs are consistent in the dataset, i.e., two connections with the same ID describe
the trajectory of the same device. We used XDR to compute trips using a stop-based method (Graells-Garrido,
Caro and Parra, 2018). This method compares the timestamp and distance between the towers associated
with two consecutive events. If the speed of the transition is above a threshold of 0.5 Km/h, the transition
encodes part of a trip. Adjacent parts of a trip are chained to identify its origin and destination towers and the
intermediary towers that serve as waypoints of each trip. Finally, we build an Origin-Destination Matrix from
all inferred trips.

To integrate this data with the other sources, we grid the origins and destinations into a regular grid defined by
the S2 Geometry system, at a level where each edge is approximately 500 metres in length. We also apply this
grid to the other sources used in this paper (see Figure 1.b) and the OD matrix built from trips (see Figure 1.c). In
addition, for each OD matrix cell, also denoted OD low, we have a set of waypoint towers that were part
of the trajectories of all trips within the corresponding flow. These waypoints are important when inferring
modes of transportation, as, in some OD pairs, there could be different routes per mode (Graells-Garrido, Caro
and Parra, 2018). We also estimate the covered distance and speed distributions of trips within a low.

The DPI data contain the number of connections in each tower to the 5,000 most accessed Internet domains
from mobile phones, representing around 80% of the all accessed domains in the country as indicated by
the mobile phone operator. These domains can be mapped to mobile phone applications or Web services, as
company or service names are encoded in domain addresses. Thus, whereas XDR indicates trajectories of
devices, DPI indicates some digital activities performed by the device owners. These activities are important, as
they may be held during transportation (Jain and Lyons, 2008). Furthermore, they may be related to the mode
of transportation the device owner is using if travelling (Graells-Garrido, Caro, Miranda, et al., 2018). In total,
we identified 887 different top-level domains, for instance, by unifying api.example.com with maps.example.com
into example.com. (see Supplementary Material Figure S2 for example application distributions, including
Uber, Spotify, Instagram, and others). Since there is a set of waypoint towers for each OD low, we estimate
a distribution of mobile phone applications accessed at each low. Some of these applications may aid in
identifying previously hard-to-detect modes, such as pedestrian and active trips (here, non-motorised) and taxi
(due to ride-hailing applications such as Uber and Cabify).

Thus, with mobile phone data, there are new distributions of OD flows in the city and the usage of mobile
phone applications in those flows. To the extent of our knowledge, this combination of XDR and DPI has not
been used in modal split problems before. Then, we expect to obtain an updated modal split by integrating
these data with official data.

**Official Data Sources**

We integrated three official data sources: a travel survey (2012), a socioeconomic characterisation survey
(2017), and a census (2017). The travel survey provides initial estimations of the mode split and other modes
of transportations distributions. The other sources are used to update the initial estimation, jointly with the
mobile phone data described earlier.

The Santiago Travel Survey is a mobility survey collected in 2012 by the Chilean Ministry of Transport and
Telecommunications (SECTRA, 2015). Since our mobile phone data were from a workday in March 2020, we
estimate an initial mode split for the municipalities understudy in a workday in morning commute hours (see
Figure 2 for the mass-transport municipal share, c.f. the income distribution of Figure 1.b, which is negatively
correlated). We also estimate the distributions of distance and speed (see Figure S3 in Supplementary Material)
per mode of transportation, as they are relevant features when discriminating the mode split in OD flows (Wang
et al., 2010). The distance ranges in the distribution were determined through a geometric progression for
distances between 0.5 and 50 Km; we observe that non-motorised predominates in short trips and that private
has a high share of trips even in arguable short distances (less than 3 Km). The speed ranges were determined
with domain experts to use discriminative ranges for all modes of transportation under consideration.

The second official source is a survey for socio-economic characterisation, named CASEN (Ministerio de
Desarrollo Social y Familia, 2018). Its 2017 edition is the last one released at the time of reporting this work
(there is an abbreviated 2020 edition that has pandemic-related questions). Relevant questions for our work
include the choice of mode of transportation to go to work, the income distribution per municipality, the household car ownership per municipality, and the distribution of mobile phone usage per municipality (either people have/do not have a mobile phone).

The third official source is the Chilean Census held in 2017 (Instituto Nacional de Estadísticas, 2019). Unfortunately, the census does not contain income data due to its abbreviated design; however, it contains questions related to educational level, which is strongly correlated with income in Chile (Garreton et al., 2020). Thus, we estimated an approximated income distribution per grid cell (see Figure 1.b for the highest income quintile geographical distribution) and distributions of age range per grid cell.

We expect these last official sources to aid the model when estimating a reconstructed modal split. On the one hand, some of these variables may have importance when estimating the mode split (such as income, household car ownership, etc.). On the other hand, other variables provide demographic controls for the model, which may help to modulate potential biases in the anonymised data (such as the distributions per age, gender, and mobile phone usage).

Data Fusion using Collective Matrix Factorisation

We integrate all the mentioned data sources into a unified dataset using a matrix-based approach. The first step is to define a coherent representation for all datasets. Then, we identify entities or concepts in all the considered datasets and their relation. Every relation can be expressed as a matrix, where rows and columns are a type of entity, and the corresponding cell value quantifies their relationship. For instance, the modal split is encoded as a matrix $M$ where municipalities are rows, and modes of transportation are columns. The matrix cell $m_{uv}$ encodes the fraction of trips that originate in municipality $u$ in mode $v$ (the matrix is row-normalised). In our context, the matrix $M$ is potentially out of date; therefore, our goal is to compute a new matrix $M'$ that contains an updated modal split by using the information present in the rest of the dataset (which is assumed to be up-to-date).
Table 1 enumerates all relationships built from the data, including the size of the matrix representation and a description of its contents. Most of the matrices are normalised or weighted as we plan to measure distributions. Only one matrix is not normalised: it is important for the model to incorporate absolute population distributions, as the digital traces come from an anonymised data source that is not representative of the population.

| Source Entity         | Target Entity          | Size            | Description                                                                 |
|-----------------------|------------------------|-----------------|-----------------------------------------------------------------------------|
| Flow between Cells     | Grid Cell (Origin)     | $34484 \times 827$ | Boolean (0 or 1) encoding if the cell (column) is the origin of the corresponding OD flow (row). |
| Flow between Cells     | Grid Cell (Destination)| $34484 \times 827$ | Boolean (0 or 1) encoding if the cell (column) is the origin of the corresponding OD flow (row). |
| Grid Cell (Origin)     | Municipality (Origin)  | $827 \times 40$  | The fraction of the population in the origin grid cell of a flow that lives within a municipality. |
| Grid Cell (Destination)| Municipality (Destination)| $827 \times 40$ | The fraction of the area of the destination grid cell of a flow that lies within a municipality. |
| Flow between Cells     | Municipality (Origin)  | $34484 \times 40$| Boolean (0 or 1) encoding if a municipality (column) is the origin of the corresponding OD flow (row). |
| Flow between Cells     | Municipality (Destination)| $34484 \times 40$| Boolean (0 or 1) encoding if a municipality (column) is the destination of the corresponding OD flow (row). |
| Municipality (Origin)  | Mode of Transportation | $40 \times 4$   | Distribution of mode of transportation usage per municipality (row-normalised). |
| Municipality (Origin)  | Mobile Phone Usage     | $40 \times 2$   | Distribution of population having a mobile phone (yes/no) per municipality (absolute values from Census). |
| Grid Cell (Origin)     | Home Income            | $827 \times 5$  | Distribution of household income quintiles per grid cell of flow origins (row-normalised). |
| Source Entity                  | Target Entity                  | Size     | Description                                                                                                                                 |
|-------------------------------|-------------------------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------|
| Grid Cell (Destination)       | Home Income                   | $827 \times 5$ | Distribution of household income quintiles per grid cell of flow destinations (row-normalised).                                            |
| Grid Cell (Origin)            | Population Age                | $827 \times 11$ | Distribution of population age ranges (in groups of 10 years) per grid cell (row-normalised).                                               |
| Home Income                   | Mode of Transportation        | $5 \times 4$ | Distribution of mode of transportation usage per household income quintile (row-normalised).                                                 |
| Municipality (Origin)         | Car Ownership                 | $40 \times 9$ | Distribution of the number of cars owned per household, per municipality (row-normalised).                                                  |
| DPI Application               | Mode of Transportation        | $887 \times 4$ | A number encoding the association between a mobile phone application (row) and a mode of transportation (column). The association is binary per mode, but the matrix is row-normalised. |
| Flow between Cells            | DPI Application               | $34484 \times 887$ | A number encoding the association between each mobile phone application and the corresponding flow. It is estimated by, first, counting the number of requests to each application in the waypoint towers of a flow; second, weighting the matrix using Log-Odds ratio; and third, replacing negative values with zeros, to ensure a positive matrix. |
| Flow between Cells            | Speed                         | $34484 \times 14$ | The distribution of speed ranges of trips within each flow (row-normalised).                                                             |
| Speed                         | Mode of Transportation        | $14 \times 4$ | The distribution of modes of transportation in each speed range (row-normalised).                                                          |
One of the characteristics of matrix factorisation methods is the ability to reconstruct matrices by means of matrix projection. Then, a matrix $M'$ could be expressed as a reconstruction of $M$ based on the other matrices available on the dataset. We can express every relation between entities $i$ and $j$ as a positive matrix $R_{ij}$ (as matrices are either based on counts or distributions, we can ensure they are positive). Previous work using factorisation has employed Non-negative Matrix Factorisation, which decomposes a matrix $R$ into the multiplication of two matrices:

$$R_{ij} \approx G_i \times G_j,$$

where $G_i$ is a latent representation of entity $i$ (rows of $R_{ij}$), and $G_j$ is a latent representation of entity $j$ (columns of $R_{ij}$). Although this approach is widely used, it is insufficient for our problem, because there is no way of incorporating other entities in the solution of the problem. In other words, both entities are clustered in latent space without considering the network of relations. The Matrix Tri-factorization approach provides a definition of a co-clustering that considers these connections (Ding et al., 2006; Žitnik and Zupan, 2014), by separating the latent representation of entities from the latent representation of their relationship:

$$R_{ij} \approx G_i S_{ij} G_j^T,$$

where $S_{ij}$ uniquely represents the relation $R_{ij}$, and $G_i$ (and, by analogy, $G_j$) represents the entity $i$ in all its relationships. Following this schema, the mode split matrix $M$ can be reconstructed from its corresponding matrix $S$, and the entity matrices $G_{\text{mode}}$ and $G_{\text{municipality}}$, which should be constructed taking into account the network of relationships.

To perform the factorisation, the following problem must be solved:

$$\min_{G \geq 0} \| R - GSG^T \|.$$

An efficient implementation based on multiplicative updates is available in the scikit-fusion library (Čopar et al., 2019).

Each entity matrix $G_i$ requires a rank parameter $k_i$. As in other factorisation methods, $k$ represents the dimension of the latent space that represents the data. If $k$ is very small in comparison to the total dimension of the entity, there is a risk of information loss; conversely, for large values of $k$, the model may overfit and give more importance to noise than to the actual information in the model. In our context, the definition of these parameters is iterative, meaning that we start with an initial set of small ranks, and then we increase them until there is no more gain regarding model goodness-of-fit.

To finish the pipeline, once the model has been adjusted, the updated mode split matrix $M'$ can be obtained directly:

$$M' = G_{\text{municipality}} S_{\text{municipality}, \text{mode}} G_{\text{mode}}^T.$$

The matrix $M'$ is the ultimate result of this pipeline. We interpret it in a case study in the next section.

### Results

Compared with the initial mode split from 2012, we observe small changes in the mode split: a slight increase in mass-transport usage from 39.44% to 40.42%, a decrease in private transport from 40.64% to 40.01%, and a
decrease in non-motorised trips from 16.15% to 15.36%; conversely, taxi trips increased from 3.76% to 4.21% (See Figure 3.a; for an analysis of parameter selection and the goodness-of-fit of this model, see Supplementary Material section “Model Goodness-of-Fit” and Figure S4). For comparison, the change between the Santiago Travel Survey between 2012 and 2001 regarding private transport was 5.1% in total (considering all trips, in contrast with our analysis that focuses only on commuting trips), and mass-transport was reduced by 6.4% (Muñoz et al., 2015). We do not observe such changes in these results. We hypothesise that some factors that affect one pattern, for instance, the decrease in availability of public transportation that may promote car ownership, is arguably negated by how migrants presumably tend to use mass-transport mainly.

Figure 3: Evolution of the modal split between 2012 and 2020 after applying the Data Fusion approach.

At the municipality level, we observe that the distribution of modal share changes had a median close to zero for all modes of transportation, with most of the changes (according to the 1.5 Interquartile range before the first quartile and after the third quartile) being situated in the range of (-16.25%, 15.47%) for mass-transport, (-17.51%, 16.57%) for private transport, (-11.07%, 10.93%) for non-motorised transport, and (-2.30%, 2.17%) for taxi (see Figure 3.b). Hence, although the modal split is seemingly similar to the initial estimation from the travel survey, there are many changes at the municipal level. To analyse these changes, we estimated the Pearson correlation coefficient \( r \) between the difference in mode share per municipality (see Table S1 in Supplementary Material). Significant correlations (after Bonferroni-correction of \( p \)-values) include the negative relationship between mass-transport and private changes (\( r = -0.78 \)), arguably the most expected relationship, as cars have increased their sales in the last years, and as these vehicles are likely to replace public transportation for their owners (Beirão and Cabral, 2007). Another negative relationship is mass-transport and taxi (\( r = -0.47 \)), showing that taxi replaces public transportation trips. This is coherent with recent findings from the literature, where ride-hailing applications tend to replace public transport and traditional taxis in Santiago, not private transportation (Tirachini and Río, 2019). We also note that taxi is positively correlated with non-motorised (\( r = 0.54 \)), hinting that areas where the circumstances favour one mode, they favour the other too.

The differences in the mode split reconstruction exhibit a geographical dependence on the distribution (see Figure 4). Most relevant changes happen in municipalities crossed or surrounded by metro or train lines. Some municipalities with a high increase in mass-transport share include Cerrillos, La Reina, and Independencia, which have access to new metro lines since 2017 (see municipality locations by name in Figure 2). On the other hand, Conchalí, in the northern area of the urban radius, which is also crossed by a new metro line, exhibited a high decrease in mass-transport share; in turn, all other three modes increased considerably. Metro operations in Conchalí started in 2019; then, we hypothesise that the social upheaval from that year delayed the expected modal shift to public transport. The three municipalities from the urban radius that decreased
Figure 4: Geographical distribution of modal split differences per municipality.
their mass-transport share the most include Lo Espejo, Lo Prado, and Macul. Lo Espejo, a disconnected and low-income municipality, had a train station inaugurated in 2017; however, it appears that the municipality shifted to private transport in a significant proportion. One plausible explanation is related to the quality of service: for Lo Espejo inhabitants, the train already comes overcrowded with people from San Bernardo. Perhaps a similar explanation can be found for Lo Prado, which has always been well-connected by public transport but whose service is of poor quality. Then, as this municipality is close to the city centre, their inhabitants shifted to taxi and non-motorised instead of private. Likewise, Macul has also decreased its share of mass-transport, although from any point in Macul, in any direction, it is possible to perform multimodal trips and connect to the metro in neighbouring municipalities. As with the other municipalities in Santiago, we argue that this difference could be explained by the poor quality of service of the public transport system, which is globally perceived as overcrowded and slow (Tirachini et al., 2017). Other municipalities that increased mass-transport are outside the urban radius, which may be explained by new services in the public transport system.

Finally, one interesting pattern for us is how the municipalities where non-motorised transport increased are primarily residential or are in non-urban areas. Further work is needed to understand these changes, as this is an aspect of transportation affected by the COVID-19 pandemic and its corresponding municipal mobility restrictions (Pappalardo et al., 2020).

Conclusion

In this work, we presented a method to update the mode split in a city by fusing data from official sources and mobile phone data. By performing a case study in a big Latin American city, we found that the proposed method generates coherent results with expectations from domain knowledge. Furthermore, the proposed method requires little data (a single day of mobile phone data, as in the case study) and supports incorporating domain knowledge or any other data source that can be expressed as a matrix. This is one of the main strengths of the proposed method, as the demand for frameworks to fuse data from a multidisciplinary perspective is increasing (Graells-Garrido et al., 2020; Meta et al., 2022).

From the specific results for Santiago, we observe that, although the overall mode split is roughly the same, there are multiple patterns of change throughout the city. This may be explained by several phenomena that affected the city simultaneously: a paramount increase in car sales, the unavailability of metro services due to the social upheaval in 2019, as well as the economic crisis caused by it, and a strong migration wave, among others. This is an opportunity to extend the model. Other official data sources about relevant phenomena could be directly integrated into the pipeline without changing how the method works, implying a transparent extension of the model and more complete outputs. We suggest this area as future work as a way to disaggregate the mode split not only by origins of trips but also by characteristics of those origins, such as the migrant population density or how much the economic crisis affected each area.

We identify two main limitations in this work. The first one is that we focused on commuting trips. This was a deliberate choice, as commuting is a daily activity that is arguably similar across different days and months within the city; however, the method could be applied at any time of the day, as there are no assumptions about commuting in the pipeline. Future extensions of this work will study the mode split at different time windows of labour and non-labour days. The second limitation is the lack of a formal evaluation with ground truth. The lack of such data was a motivation of this work, and although we followed traditional machine learning methods to select model parameters, there is still a need to evaluate this type of contribution.

Pressing challenges will affect cities at even larger scales than in recent years. Urban planners and policymakers will need cost-effective tools to work at finer resolutions than their actual input sources. This will require the collaboration of scientists, private and public institutions, and the development of methods such as the one proposed in this work.

Availability of data. The Telefónica Movistar mobile phone records have been obtained directly from the mobile phone operator through an agreement between the Data Science Institute from Universidad del Desarrollo 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 for researchers who meet the criteria for access to confidential data. The other datasets used in this study are publicly available.

Acknowledgments. We thank GPA Movistar for facilitating the data for this study, and Leo Ferres for insightful discussion. Daniela Opitz was partially funded by ANID #PAI77190057.

References

Acar C and Dincer I (2020) The potential role of hydrogen as a sustainable transportation fuel to combat global warming. *International Journal of Hydrogen Energy* 45(5). Elsevier: 3396–3406.

Allen SK, Barros V, Burton I, et al. (2012) Summary for policymakers. In: Field CB, Barros V, Stocker TF, et al. (eds) *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, pp. 3–22. DOI: 10.1017/CBO9781139177245.003.

Antoniou C, Balakrishna R and Koutsopoulos HN (2011) A synthesis of emerging data collection technologies and their impact on traffic management applications. *European Transport Research Review* 3(3). Springer: 139–148.

Bachir D, Khodabandelou G, Gauthier V, et al. (2019) Inferring dynamic origin-destination flows by transport mode using mobile phone data. *Transportation Research Part C: Emerging Technologies* 101: 254–275. DOI: https://doi.org/10.1016/j.trc.2019.02.013.

Bantis T and Haworth J (2017) Who you are is how you travel: A framework for transportation mode detection using individual and environmental characteristics. *Transportation Research Part C: Emerging Technologies* 80. Elsevier: 286–309.

Beirão G and Cabral JS (2007) Understanding attitudes towards public transport and private car: A qualitative study. *Transport policy* 14(6). Elsevier: 478–489.

Brueckner JK (2005) Transport subsidies, system choice, and urban sprawl. *Regional Science and Urban Economics* 35(6). Elsevier: 715–733.

Colvile R, Hutchinson EJ, Mindell J, et al. (2001) The transport sector as a source of air pollution. *Atmospheric environment* 35(9). Elsevier: 1537–1565.

Čopar A, Zupan B and Zitnik M (2019) Fast optimization of non-negative matrix tri-factorization. *PloS one* 14(6). Elsevier: 1537–1565.

Čopar A, Zupan B and Zitnik M (2019) Fast optimization of non-negative matrix tri-factorization. *PloS one* 14(6). Public Library of Science San Francisco, CA USA: e0217994.

Dabiri S and Heaslip K (2018) Inferring transportation modes from GPS trajectories using a convolutional neural network. *Transportation research part C: emerging technologies* 86. Elsevier: 360–371.

Ding C, He X and Simon HD (2005) On the equivalence of nonnegative matrix factorization and spectral clustering. In: *Proceedings of the 2005 SIAM international conference on data mining*, 2005, pp. 606–610. SIAM.

Ding C, Li T, Peng W, et al. (2006) Orthogonal nonnegative matrix T-factorizations for clustering. In: *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*, New York, NY, USA, 2006, pp. 126–135. KDD '06. Association for Computing Machinery. DOI: 10.1145/1150402.1150420.

El Faouzi N-E, Leung H and Kurian A (2011) Data fusion in intelligent transportation systems: Progress and challenges—a survey. *Information Fusion* 12(1). Elsevier: 4–10.

Garreton M, Basauri A and Valenzuela L (2020) Exploring the correlation between city size and residential segregation: Comparing Chilean cities with spatially unbiased indexes. *Environment and Urbanization* 32(2). SAGE Publications Sage UK: London, England: 569–588.

González MC (2019) Mining urban lifestyles: Urban computing, human behavior and recommender systems. Computing. Institution of Engineering; Technology. DOI: 10.1049/PBPC035G_ch5.

Graells-Garrido E, Peredo O and García J (2016) Sensing urban patterns with antenna mappings: The case of Santiago, Chile. *Sensors* 16(7). Multidisciplinary Digital Publishing Institute: 1098.

Graells-Garrido E, Caro D and Parra D (2018) Inferring modes of transportation using mobile phone data. *EPJ Data Science* 7(1). Springer Berlin Heidelberg: 49.

Graells-Garrido E, Caro D, Miranda O, et al. (2018) The WWW (and an H) of mobile application usage in the city: The what, where, when, and how. In: *Companion proceedings of the the web conference 2018*, 2018, pp. 1221–1229.

Graells-Garrido E, Peña-Araya V and Bravo L (2020) Adoption-driven data science for transportation planning:
Methodology, case study, and lessons learned. *Sustainability* 12(15). DOI: 10.3390/su12156001.

Huang H, Cheng Y and Weibel R (2019) Transport mode detection based on mobile phone network data: A systematic review. *Transportation Research Part C: Emerging Technologies* 101. Elsevier: 297–312.

Instituto Nacional de Estadísticas (2019) Censo 2017. Available at: http://www.censo2017.cl.

Jain J and Lyons G (2008) The gift of travel time. *Journal of transport geography* 16(2). Elsevier: 81–89.

Kuang D, Ding C and Park H (2012) Symmetric nonnegative matrix factorization for graph clustering. In: *Proceedings of the 2012 SIAM International Conference on Data Mining*, pp. 106–117. DOI: 10.1137/1.9781611972825.10.

Li Y, Wu F-X and Ngom A (2018) A review on machine learning principles for multi-view biological data integration. *Briefings in bioinformatics* 19(2). Oxford University Press: 325–340.

Meta I, Cucchietti FM, Navarro-Mateu D, et al. (2022) A physiology-inspired framework for holistic city simulations. *Cities* 126: 103553. DOI: https://doi.org/10.1016/j.cities.2021.103553.

Ministerio de Desarrollo Social y Familia (2018) Encuesta CASEN 2017. Available at: http://observatorio.ministeriodesarrollosocial.gob.cl/encuesta-casen-2017.

Muñoz V, Thomas A, Navarrete C, et al. (2015) Encuesta origen-destino de Santiago 2012: Resultados y validaciones. *Estudios de Transporte* 19(1).

Pappalardo L, Cornacchia G, Navarro V, et al. (2020) A dataset to assess mobility changes in Chile following local quarantines. *arXiv preprint arXiv:2011.12162*.

Rodríguez-Núñez E and García-Palomas JC (2014) Measuring the vulnerability of public transport networks. *Journal of transport geography* 35. Elsevier: 50–63.

Rowe F (2022) Big data and human geography. In: Demeritt D and Lees L (eds) *Concise Encyclopedia of Human Geography. Encyclopedias in the Social Sciences Series*. Edward Elgar, p. xx.

SECTRA (2015) Informe ejecutivo encuesta origen-destino de viajes Santiago. Secretaría de Transportes, Ministerio de Transporte y Telecomunicaciones.

Somma NM, Bargsted M, Disi Pavlic R, et al. (2021) No water in the oasis: The Chilean Spring of 2019–2020. *Social Movement Studies* 20(4). Taylor & Francis: 495–502.

Thomson JM (1998) Reflections on the economics of traffic congestion. *Journal of transport economics and policy*. JSTOR: 93–112.

Tirachini A and Río M del (2019) Ride-hailing in Santiago de Chile: Users’ characterisation and effects on travel behaviour. *Transport Policy* 82. Elsevier: 46–57.

Tirachini A, Hurtubia R, Dekker T, et al. (2017) Estimation of crowding discomfort in public transport: Results from Santiago de Chile. *Transportation Research Part A: Policy and Practice* 103. Elsevier: 311–326.

Urquieta Ch. C (2019) Dura pérdida para el Metro: No tiene seguros para estaciones ni trenes. *CIPER*. Available at: https://www.ciperchile.cl/2019/10/24/dura-perdida-para-el-metro-no-tiene-seguros-para-estaciones-ni-trenes/ (accessed 15 April 2022).

Wang H, Calabrese F, Di Lorenzo G, et al. (2010) Transportation mode inference from anonymized and aggregated mobile phone call detail records. In: *Intelligent transportation systems (ITSC), 2010 13th international IEEE conference on*, 2010, pp. 318–323. IEEE.

Zannat KE and Choudhury CF (2019) Emerging big data sources for public transport planning: A systematic review on current state of art and future research directions. *Journal of the Indian Institute of Science* 99(4). Springer: 601–619.

Žitnik M and Zupan B (2014) Data fusion by matrix factorization. *IEEE transactions on pattern analysis and machine intelligence* 37(1). IEEE: 41–53.
Supporting Information

Paper Overview

Figure S1: Schematic view of the methods and data used to estimate an updated mode split for a city. For instance, the mode split from 2012 can be updated using official data from 2017 and mobile phone data from 2020.
Figure S2: Distribution of mobile application usage (from DPI data) in towers in towers. Each tower is depicted with a circle with area proportional to the intensity of the application usage. a) Instagram (photo-sharing social network). b) Uber (ride-hailing). c) Waze (GPS routing). d) WhatsApp (messaging). e) Spotify (audio). f) Niantic Labs (augmented reality games such as Pokémon Go). Black lines represent primary streets in the city, cyan lines represent the metro network. Urban network data has copyright from OpenStreetMap contributors, used with permission.
Distance and Speed Distributions from Travel Survey

Figure S3: Mode split per distance and speed of trips, estimate from the Santiago Travel Survey 2012.

Correlations in Mode Shift

Table S1. Pearson correlation coefficients between changes in the mode split per mode of transportation. All p-values are Bonferroni-corrected.

| Modes                        | Pearson r | p-value |
|------------------------------|-----------|---------|
| mass-transport, non-motorised| -0.41     | > 0.05  |
| mass-transport, taxi         | -0.47     | < 0.05  |
| mass-transport, private      | -0.78     | < 0.001 |
| non-motorised, taxi          | 0.54      | < 0.01  |
| non-motorised, private       | -0.23     | > 0.05  |
| taxi, private                | 0.02      | > 0.05  |

Model Parameterization

Table S2. Entity Ranks.

| Entity                      | Matrix G Rank |
|-----------------------------|---------------|
| Mobile Phone Usage          | 2             |
| Mobile Phone Tower          | 2             |
| Mode of Transportation      | 4             |
| Home Income                 | 4             |
| Population Age              | 8             |
| Car Ownership               | 8             |
| Speed                       | 8             |
| Distance                    | 8             |
| Municipality (Origin)       | 32            |
| Municipality (Destination)  | 32            |
| Grid Cell (Origin)          | 512           |
| Grid Cell (Destination)     | 512           |
Model Goodness-of-Fit

Each entity required a rank-parameter $k$. The matrix rank for each entity was determined through iteration (see Table S2 for final values). We started with a rank of 2 for all matrices, and at every iteration we increased the rank of the entities involved in the relation in powers of two, unless the rank was higher than the cardinality of the corresponding entity. For every iteration, we estimated the mean reconstruction error, where the reconstruction error of one matrix is the norm of the difference between its reconstruction and the original, divided by the norm of the original matrix. We used the elbow method as a stop criterion. This is, we stopped when the mean reconstruction error did not improve from one iteration to another (see Figure S4.a). Some of the largest matrices and the most heterogeneous from the data encoded population behaviour showed a high reconstruction error (up to 0.96, see Figure S4.b). We cannot expect to reconstruct every detail; instead, we aim to find the most important patterns connected to the other matrices in the data. Another matrix with high error is the relationship between application usage and mode of transportation, which has a reconstruction error of 0.55. This value was expected to be high, as few applications were known to be associated with modes of transportation. We assumed that other applications were not associated with a mode, and thus, the matrix was filled with a majority of zeroes. In the reconstruction, the model found a non-null association between other applications and modes of transportation. Conversely, matrices related to population statistics had errors that lie between very small (0.03, absolute population values with respect to mobile phone usage) to moderate (0.19, age distribution per grid cell).

![Figure S4: Analysis of model error. a) A line chart of the mean reconstruction error reduction at every iteration. In a new iteration the entity ranks were increased in powers of two. b) The reconstruction error for each relation in the integrated dataset.](image)