Issue 20(4), 2020, pp. 269-285
https://doi.org/10.18757/ejtir.2020.20.4.5429

TU Delft OPEN

EJTIR

ISSN: 1567-7141
http://ejtir.tudelft.nl/

Modelling cellphone trace travel mode with neural networks using transit smartcard and home interview survey data

James Vaughan[1]
University of Toronto Transportation Research Institute, Canada.

Ahmadreza Faghih Imani[2]
Centre for Transport Studies, Imperial College London, United Kingdom.

Bilal Yusuf[3]
Metrolinx, Canada.

Eric J. Miller[4]
University of Toronto Transportation Research Institute, Canada.

This study proposes a framework to impute travel mode for trips identified from cellphone traces by developing a deep neural network model. In our framework, we use the trips from a home interview survey and transit smartcard data, for which the travel mode is known, to create a set of artificial pseudo-cellphone traces. The generated artificial pseudo-cellphone traces with known mode are then used to train a deep neural network classifier. We further apply the trained model to infer travel modes for the cellphone traces from cellular network data. The empirical case study region is Montevideo, Uruguay, where high-quality data are available for all three types of data used in the analysis: a large dataset of cellphone traces, a large dataset of public transit smartcard transactions, and a small household travel survey. The results can be used to create an enhanced representation of origin-destination trip-making in the region by time of day and travel mode.

Keywords: *cellphone traces, cellular network data, transit smartcard, machine learning, mode detection.*

1. Introduction

In recent years there has been a growing interest in incorporating different forms of emerging passive data sources into analysis of transportation systems and travel behaviour. Cellphone traces

[1] A: 35 St. George Street, Toronto, Canada T: +1 416 978 4076 F: +1 416 978 6813 E: james.vaughan@utoronto.ca
[2] A: South Kensington Campus, London, SW7 2AZ, United Kingdom E: s.faghih-imani@imperial.ac.uk
[3] A: 97 Front St West, Toronto, Canada, E: bilal.yusuf@metrolinx.com
[4] A: 35 St. George Street, Toronto, Canada T: +1 416 978 4076 F: +1 416 978 6813 E: eric.miller@utoronto.ca

collected from cellular network base transceiver stations (BTSs) provide a very large sample of data with widespread coverage and low cost for a long period, especially for urban regions due to the increasing market penetration of cellphones. Given the ubiquity of cellphones in recent years, imputing people's movement using BTS cellular network data would generate a representative travel pattern for the population in a region, with relatively good temporal coverage at a low cost. A cellular network typically records several cellphone-to-infrastructure connection events namely handovers (HO), location updates (LU) and call detail records (CDR). HO and CDR provide data of communication events such as calls or messages and LU notify the cellular network when a cellphone moves from one BTS to another.

One benefit of using cellular network data for transportation analysis is the ability to use existing infrastructure. It is not required to install new antennas or devices, in contrast to other technologies such as Bluetooth. Unlike the methods using GPS services, for using cellular network data, there is no need to ask people to install an app on their smartphones. All cellphone users' movements can be potentially recorded using the information collected from BTS antennas. This also results in a major privacy concern; it is not possible for users to opt-out or to refuse to participate in the data collection as long as they use their cellphones. Another issue with this type of data is that individual characteristics such as socio-demographic attributes are unknown. Therefore, if the population of cell phone users are not representative of the entire population, it is not possible to correct for the sample bias. Additionally, if a person carries two or more cellphones, there is no way to identify that person and just count the movements once.

Each BTS serves a certain area, usually referred to as a BTS cell. The spatial precision of the cellular network data depends on the size of this cell. While in dense urban areas the cell sizes are usually small, the cell sizes in rural areas can reach several kilometers. Also, in rural areas, the service coverage is not usually complete, which can potentially generate gaps in people's travel patterns collected from the cellular network data. This lack of spatial precision increases the level of uncertainty and makes the cellular network data more suitable for aggregate traffic measures such as generating OD matrices (Caceres et al. 2013; Gundlegård et al. 2016; Xu et al. 2017), estimating road congestion (Caceres et al. 2012; Gundlegård et al. 2016; Toole et al. 2015) or investigating long-distance travel (Bekhor et al. 2013; Janzen et al. 2018).

While cellphone traces can be used to identify mobility patterns, the lack of trip information such as mode and purpose is a key limitation for such data to replace traditional travel surveys, as well as in comparison to many app-based methods. Mode and purpose information are two essential components of any transportation study. Imputation of mode and purpose is further complicated by technical limits to the spatial and temporal precision of the cellphone traces that can generally be achieved. While several studies have tried to identify trip purposes (Alexander et al. 2015; Xu et al. 2017; Yin et al. 2017), to date, the question of mode imputation has often been ignored, with researchers mostly focussing on auto-based travel (Alexander et al. 2015; Gundlegård et al. 2016; Toole et al. 2015; Yin et al. 2017). Additional processing and imputation are necessary to detect travel mode for trips identified from cellphone traces.

This study aims to bridge this gap in the literature by developing a deep neural network model that identifies mode of travel for trips identified from cellphone traces. The empirical case study region is Montevideo, Uruguay, where high-quality data are available for all three types of data used in the analysis: a large dataset of cellphone traces, a large dataset of public transit smartcard transactions, and a small household travel survey. In the household travel survey and public transit smartcard transaction datasets, the travel modes, origin and destination locations and start- and end-times are known and reported for every trip. Thus, it is possible to create a set of artificial pseudo-cellphone traces for those trips with the mode of travel known. The generated artificial pseudo-cellphone traces then can be used to train a deep neural network classifier. In this study, the modes considered are car, public transit and active transportation (walk and bicycle combined). With the exception of the Bachir et al. (2019) study (discussed further below), the use of travel

survey and transit smartcard data to impute cellphone trace travel modes does not seem to have been explored in the literature. Also, to the best of the authors' knowledge, the proposed "pseudo-trace" method for training the mode imputation procedure has not been previously tested. Thus, a primary purpose of this paper is to explore the viability of this approach.

The rest of the paper is organized as follows. Section 2 provides a review of literature on use of cellular network data in transportation analysis. Section 3 presents the data used in our analysis. Our proposed method for imputing travel mode of cellphone traces is explained in Section 4 and the results for the Montevideo region are presented in Section 5. We conclude the paper in Section 6.

2. Literature Review

With the advance of technology in recent years, investigating people movements using a carry-on device has become popular due to the potentials to obtain accurate data for a large sample with relatively low cost (Lee et al. 2016). Earlier studies have examined the use of cellular network data for transportation planning purposes from various dimensions. Several studies have reviewed current methods and practices, potentials and limitations of using cellular network data for transportation planning analyses (Caceres et al. 2008; Chen et al. 2016; Çolak et al. 2015; Huang et al. 2019; Jiang et al. 2013; Wang et al. 2018). Studies investigating cellular network data to better understand human mobility patterns typically focus on identifying activities, trips, and spatial-temporal variations in travel patterns (Becker et al. 2011; Bekhor and Shem-Tov 2015; Pucci et al. 2015; Xu et al. 2017; Zahedi and Shafahi 2017). Specifically, detecting home locations is considered important to yield useful insights into people's travel patterns. Although passive cell phone location data can be sparse in both space and time, inferred home locations are found to be relatively accurate as they are locations which are repetitively visited. The identified home locations further are used as anchors to examine other activities and trips (Ahas et al. 2010; Xu et al. 2015).

Another typical use of cellular network data in the literature is estimating origin-destination (OD) matrices. However, since the BTS cells are not generally aligned with traffic analysis zones or census tracts and must be re-aggregated or redistributed, the level of accuracy decreases for transportation analysis. Nevertheless, it is possible to obtain BTS cell-to-cell OD matrices from cellular network data. In the transportation literature, several studies have focused on extracting OD matrices from cellular network data for different regions around the world (Caceres et al. 2013; Caceres et al. 2007; Calabrese et al. 2011; Demissie et al. 2016; Iqbal et al. 2014; Larijani et al. 2015; Mellegard et al. 2011; Nanni et al. 2014; Pucci et al. 2015; Zhang et al. 2010). The extracted OD matrices are then used for different purposes, such as optimizing public transport network service (Berlingerio et al. 2013) or estimating traffic flows (Gundlegård et al. 2016). Several recent studies have employed cellular network data to develop activity-based travel demand models (Alexander et al. 2015; Pozdnoukhov et al. 2016; Yin et al. 2017). Given the level of spatial precision, several recent studies suggested that cellular network data is well suited for evaluating long-distance travel patterns and obtaining large samples of data for such trips (Bekhor et al. 2013; Janzen et al. 2018; Janzen and Axhausen 2015).

In the field of traffic engineering, cellular network data are also used to estimate traffic flows and to derive routes traveled. In urban areas where the BTS cell sizes are relatively small, it is possible to assign movements to the road network. Earlier studies have worked on estimating volume and speed, improving traffic assignments and inferring real-time traffic information using cellular network data (Bar-Gera 2007; Caceres et al. 2012; Janecek et al. 2015; Järv et al. 2012; Tettamanti and Varga 2014; Toole et al. 2015; Wu et al. 2015). Due to uncertainty caused by the larger BTS cell size and the possibility of having several roads within a cell for the movements' direction, most of these studies looked at aggregated measures of traffic. Further, cellular network data can provide middle

points along the way between origins and destinations and thus has also been used for route choice analysis.

Traditionally, many studies have focused on identifying mode of travel from GPS data traces (Bohte and Maat 2009; Broach et al. 2019; Semanjski et al. 2017). Since cellular network data typically have less spatial and temporal accuracy than GPS data, mode imputation is more challenging and requires more complex processing methods (Huang et al. 2019). In recent years, there has been several studies exploring methods to impute travel mode for trips generated from cellular network data. Several of these studies focused on intercity trips and therefore travel modes are generally categorized as flight and ground modes, train or highway (Doyle et al. 2011; Hui et al. 2017; Schlaich et al. 2010; Smoreda et al. 2013).

Few studies have tried to impute travel mode for urban trips using cellular network data. Phithakkitnukoon et al., 2017 used one year of CDR data in Portugal and estimate the probability of commutes made by car or transit by considering the proximity of cell towers and suggested route alternatives from Google Map between the origin and destination of the trip (Phithakkitnukoon et al. 2017). Their study focused on trips made between home and work locations and estimated the mode share of these trips for each user, with promising results in terms of aggregated mode shares when compared with travel survey data. Similarly, Qu et al., 2015 employed transportation network data and three weeks of CDR data for about 2 million cellphone users in Boston to estimate transportation mode shares in each traffic zone instead of transportation mode of each trip (Qu et al. 2015).

Wang et al, 2010 proposed a framework to infer travel mode (drive and public transport) from CDR data based on an unsupervised machine learning clustering method on trip duration and Google Map travel times (Wang et al. 2010). Xu et al, 2011, used a Hidden Markov Model to probabilistically identify travel mode for cellphone users in a small city in China. For about 500 individuals, the travel modes were known and labelled as the information were collected using a questionnaire survey. The method was based on the observed average speed and the modes' speed distribution. The modes considered were drive, bike and walk (Xu et al. 2011). Sohn et al., collected one month of cellular network data for three users who at the same time recorded their travel dairies (Sohn et al. 2006). Using that database, they were able to train a two-stage classification scheme: first stage classifying between stationary and non-stationary events and the second stage determining non-stationary events as walking or driving. Their analysis resulted in a prediction accuracy of 85%, indicating that the trained classifiers would reasonably work if the travel mode is known for trips generated by cellular network data.

Two recent studies have proposed methods to infer transportation modes using large-scale cellular network data (Bachir et al. 2019; Graells-Garrido et al. 2018). Graells-Garrido et al., 2018, developed an algorithmic pipeline to assign mode probabilities to BTS towers and use them to infer mode distribution of commuting in the city of Santiago, Chile. In their analysis, a set of modal clusters for BTS towers were generated based on urban/transportation infrastructure (GTFS and OpenStreetMap data). Depending on the BTS clusters within the commute of each user, the probability of mode of travel is then assigned. Similarly, Bachir et al. (2019) constructed OD matrices for the Greater Paris area processing two months of cellular network data. They assumed a zone for each BTS service area and, based on the transport network and usage in that zone, they clustered cellular network zones. Then, they derived transport mode probabilities per cluster using a Bayesian inference method considering the prior transport mode probability from the travel

survey for that zone. They then evaluated the inferred modal OD matrices with travel survey and transit smartcard datasets[5].

In this study, we further contribute to the growing literature on travel mode inference of cellular network data by developing a deep neural network[6] model for trips identified from a large dataset of cellphone traces in Montevideo, Uruguay. Compared to earlier studies, except (Bachir et al. 2019; Graells-Garrido et al. 2018), the cellphone traces dataset in our analysis include all cellphone to BTS connection events and are not limited to only CDR events. Given the region's mode distribution, we consider car, bus and active transportation as possible modes of travel. Rather than looking at the probability of the modes at a BTS zone level, we examine cellphone traces at a daily level and infer the mode probabilities for each trip. In our study for the training dataset we create artificial pseudo-cellphone traces from a household travel survey which has travel mode, origin and destination locations, and start-time and end-time reported for every trip. In addition, for transit trips, we augmented the dataset with pseudo-cellphone traces generated from public transit smartcard transactions. We train a deep neural network model on this labelled dataset and apply the trained model to infer modes of travel for the cellphone traces from cellular network data.

3. Data

Several data sources are used in this study including cellular network data, household travel survey data, smartcard transit transaction data as well as census data and road and transit networks. These datasets are briefly described in this section.

The primary dataset in this study comes from Antel, the largest telecom company in Uruguay7. It includes minute-by-minute cellphone traces for the month of May, 2018 of a random sample of 40% of Antel cellphone users within Montevideo and the surrounding metropolitan area. Overall, the dataset includes more than 117,862,000 cellphone traces for about 948,600 unique cellphones. These cellphone traces were preprocessed from raw data by Antel to eliminate as much noise in the data as possible and to preserve user anonymity. The geolocation data from Antel include CDR events, handovers and location updates. Trace data are temporally reported in minutes, thereby providing excellent temporal precision. The data is spatially aggregated to a 135 traffic zone system for the region. The average area of the zones in the study area is about 37.1 km2, with the smallest zone having an area of 0.38 km2 and the largest zone about 1,020 km2.

The Montevideo Household Mobility Survey (MHMS) is a very high-quality one-day household travel survey, but with a very small sample (0.34%), conducted in August, 2016. It is a classic home interview survey in which trained interviewers survey randomly selected households in their homes. Despite the small size, the survey is deemed to be a representative sample of trip-making in Montevideo (Miller et al. 2017). The dataset consists of 2,230 households and 5,946 individuals' interviews, reporting 12,546 trips for a single weekday. For every trip reported, details of every stage of the trip are also reported. This stage-level information allows us to create a dataset comparable with cellphone traces and public transport smartcard transaction datasets. The average age of individuals in the survey is 38.8 years; 53.1% of them are female. The MHMS data is spatially aggregated to census segments. Census segments are groups of blocks (usually between 6 to 12

[5] In recent years, transit smart card data have been extensively used in various transport studies and travel behaviour analysis (for examples, see (Axhausen 2017; Chapleau et al. 2018; Ordóñez Medina 2018; Pelletier et al. 2011; Zhang et al. 2018)).

[6] Neural network models and other machine learning methods have recently been widely used in transportation literature (see (Hillel et al. 2019) for a review of machine learning methods for modelling travel mode choice, and see (Koushik et al. 2020) for a review of machine learning methods for activity-travel behaviour analysis).

[7] Antel has about 60% of the cellular market share in Uruguay, at the time of this study.

blocks) that are the spatial units used by the 2011 Uruguayan census. There are about 1,720 census segments in the Montevideo region with an average area of 2.91 km2. The mode share of the reported trips in the MHMS dataset is presented in Table 1. Walking has the largest mode share, given that over a third of the trips recorded are short. Combining drives and passengers, cars account for 29.2% of trips, while the bus system carries about 25.2% of the trips.

Table 1 MHMS mode share

Travel Mode	Mode Share in %	Number of Trips
Walk	34.0	4265
Bike	3.5	439
Auto Passenger	10.0	1251
Auto Driver	19.2	2410
Motorcycle	6.1	769
Bus	25.2	3166
Other	2.0	246

MHMS origin-destination (O-D) trips by mode and time of day are compared with the O-D matrices generated by the processed cellphone data in (Faghih-Imani et al. 2018). In general, the spatial-temporal distributions of the two datasets appear to be comparable.

All public transit smartcard transactions (representing 82.5% of all transit trips within the Montevideo region) were provided by the Intendencia de Montevideo for the same May, 2018 time period and traffic zone system. These consisted of 29,868,716 recorded transactions from 734,569 unique smartcards. The transit system is tap-on only, thus requires additional processing to identify alighting stops. The Intendencia in-house algorithm successfully identified 62% of these alighting locations, which were augmented to the transaction records. Unique (but anonymized) identifiers were attached to smartcard records so that the trip-making behaviour of the smartcard owner could be tracked day-to-day within the observation period.

A computerized Montevideo road network was obtained from the OpenStreetMap database. Further, the data regarding the public bus system including routes, stops and frequencies are gathered from the open data portal of Montevideo government. These two networks were imported into Emme software and used in our traffic assignment model to estimate link travel times, volumes and congestion in the network, along with O-D travel times. The road and bus networks are presented in Figure 1. Finally, in our analysis, the latest census data (2011) are used to compute the population, household and job density variables.

Figure 1. Road and Bus Network in Montevideo

4. Method

For our analysis, an assumption is made that any time a cellphone user stays more than 30 minutes in a zone indicates participation in an activity at that location (similar assumptions for a minimum stationary duration to identify activities has been made in earlier literature, for examples see (Bonnel et al. 2015; Chen et al. 2014; Schlaich et al. 2010; Zahedi and Shafahi 2017)). With this assumption, 32,417,002 activities are identified for the four weeks of data. Further, trips are defined when a user moves between two activities, with first activity location identified as the trip origin and the second activity location as the trip destination. More than 14 million trips are identified in within the dataset. The average estimated daily trip rate per person is 1.99, with a minimum of 1 and maximum of 13.54. This daily trip rate is slightly lower than that observed in the 2016 MHMS of 2.11, which may be due to the fact that intrazonal trips cannot be identified from cellular network data[8]. The next step is to assign travel mode for these identified trips.

A deep neural network approach was adopted in this study as a logical approach to imputing cellphone trace travel modes for the large datasets being used. The key concern in developing any neural network is the availability of a "labelled" dataset containing observed outcomes (in this case, travel mode) which can be used to train and validate the neural network before it as applied to an application dataset. Given that cellphone traces by definition do not provide travel mode information, the key methodological question is how to construct a labelled training set in this application.

Figure 2 displays the overall approach developed in this study to imputing trip mode for the cellphone traces. The first, key (and novel) step in this procedure is to convert MHMS and transit smartcard OD trips into pseudo cellphone traces; i.e., for each trip, convert it into a "trace" at the same level of spatial and temporal aggregation as the actual Antel traces. These pseudo-traces were then used as labelled input data to train the neural network model.

In order to construct these traces, the MHMS OD trips must first be assigned to paths (routes) through the road networks (depending on each trip's chosen mode). Similarly, transit OD trips from the smartcard transaction dataset (only those with estimated alighting locations are considered) have been assigned to the transit, given that the transfer stops are known for trips with transfers. To do this, road and transit networks for Montevideo were constructed within the Emme network modelling software system. Maximum utility paths through the road and transit networks were found for auto (drive, passenger, taxi, motorcycle) and public transit trips, respectively. Active mode traces were constructed by taking the shortest distance paths through the road network at an assumed speed of 4 kph[9].

In this model, the day is divided into five-minute segments. A trip is defined by:

1. Whether it is occurring during a given five-minute segment (=1) or not (=0); i.e., whether the person is moving during this time segment.
2. The distance travelled during the five-minute segment.

[8] The average MHMS daily trip rate of 2.11 may also be low relative to that observed in some other urban regions. Depending on details of the survey method, household travel surveys can be subject to some under-reporting of short trips, which respondents may tend to fail to report. The authors have no independent data to assess this under-reporting hypothesis in this case. In any event, the analysis in this paper is not focussed on estimating trip rates, but rather the trip modes of the identified trips, and so this concern is of second-order importance at best. To re-iterate, it is likely that the cellphone trace data is under-estimating trip rates somewhat due to the intrazonal limitation discussed, but this should not affect the travel mode imputation results which are the focus of this paper.

[9] Since it was not possible to calibrate Montevideo-specific assignment model parameters within this study, parameters from Toronto's GTAModel V4.0 were used.

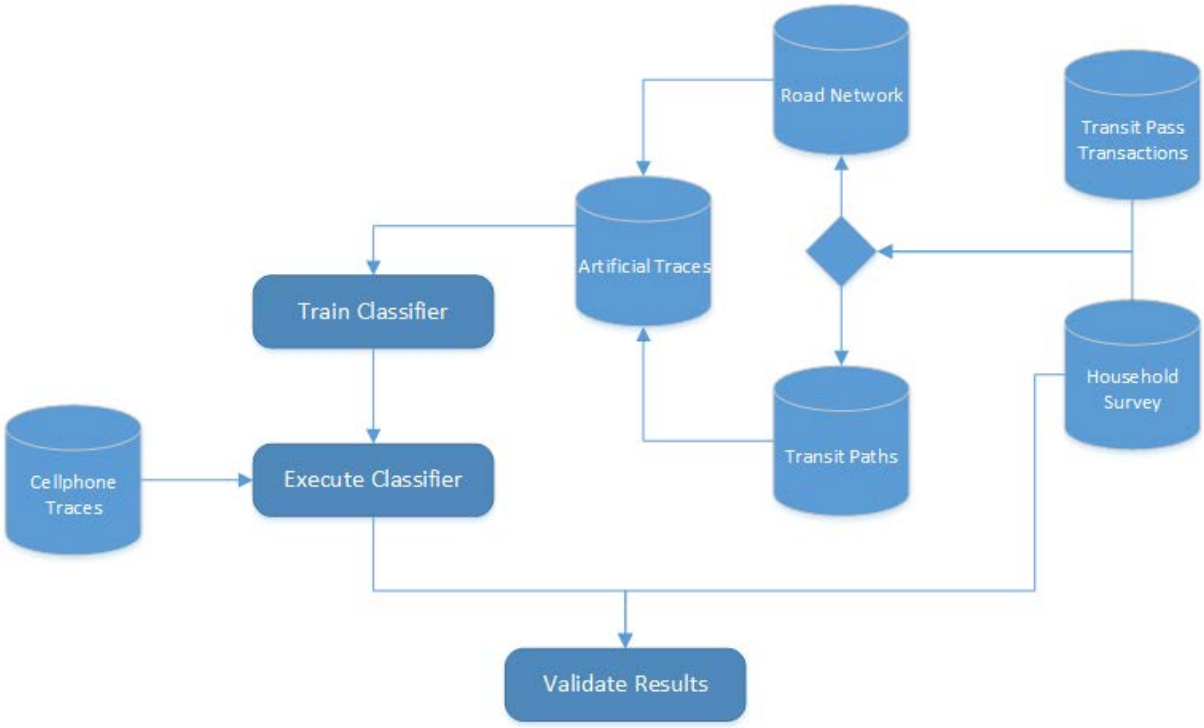

Figure 2. Analysis Approach

In addition to these two features, total trip distance and land use attributes at each trip's origin and destination are considered. The job of the neural net model is to determine that probability of each trip being made by the auto, transit or active modes, given the time of day, trip length (in time and distance) and distances travelled per time segment (approximate speed) for the trip, and origin and destination land use attributes: population, home and job density variables.

The neural net classifier model developed in this study consists of three hidden layers, with 500 neurons per layer using the linear rectifier activation function. Weights were randomly initialized from a normal distribution. Backpropagation combined with Adaptive Moment Estimation (Adam) was used to update the weights in each step of the training session (Kingma and Ba 2014). Weights were chosen to maximize the cross-entropy (effectively a log-likelihood) function. A softmax activation function was used for the output layer, in order to generate probabilities to assign to the three modes. Note that the softmax activation function is effectively a logit model which assigns for each trace a probability for each travel mode.

Three modes were imputed for cellphone trips in this analysis:

- ***Auto***: Any trip made in a passenger car or equivalent (e.g., a light truck). This includes auto-drivers and auto passengers in private cars and taxi/Uber passengers[10].
- ***Public transit***: Any trip using the STM (Sistema de Transporte Metropolitano) bus system.
- ***Active travel***: Any trip all-way from origin to destination by walking or bicycle modes. Given the spatial precision of the cellphone data it is not possible to distinguish between walk and bicycle trips with any confidence.

[10] Note that some double-counting may occur in the dataset, given that some taxi/Uber drivers may also be included in the dataset. Their trips while carrying passengers should not be included in the trip count. It may be possible to identify these drivers in the dataset and delete them from the mode imputation analysis, but this has not been attempted in this study.

Since, we do not observe intrazonal trips in cellphone traces, only interzonal trips from the MHMS and smartcard transaction datasets are considered.

The pseudo-traces were split into two datasets, 80% of trips for estimation and the rest for testing. Since the smartcard transit trips are more accurate in terms of time and transfer stops, we substituted the transit trips of the MHMS data with appropriate observed trips from the smartcard transaction data. 5000 additional smartcard transit trips were added to the estimation set to enhance the richness of the transit trips used for training; 1000 smartcard trips were similarly added to the test set. The choice of 5000 as the number to add was based on the need to keep the estimation dataset within computer memory limits. To balance the dataset, non-transit mode records from MHMS were duplicated. Adding the Smartcard data significantly improved the estimation as the observed stop location information is more accurate than the transit path assignment generated from EMME for the MHMS transit trips. Using the transit smartcard records improved transit mode recall from 67.1% to 87.5% and transit mode precision from 74.8% to 93.0% in the validation tests.

The cellphone trace data also need to be processed into features for the neural network. For every cellphone user, each person's day is extracted and trips made in a day are filtered. For trips, if no middle zone is observed in the trip, the distance between the two zones was recorded at the time bin between the start and end time. If a middle zone is recorded in a trip, the distance from the previous zone (origin if the first) is recorded at the time when the middle zone was detected. This way, for all trips identified in cellphone trace dataset, similar features of the trained model are generated. The trained neural network model is then applied to the cellphone traces to infer the probability of travel modes.

5. Results

Several neural network specifications have been tested. A model with three hidden layers and 500 neurons per layer was found to provide the best performance. Overall, 19,499 trips with pseudo-traces were used to train the neural net model and 2,106 trips with pseudo-traces were used to test the trained model. The predictions are broken down into two statistics, recall and precision for each mode. The recall of a mode is the probability that the mode is predicted given the mode was observed. Precision on the other hand is the probability that when a mode is selected that it was the observed mode. An excellent fit of 87% correct predictions by the trained model on its training set was achieved. When the trained model was applied to the test set a similarly excellent fit of 84.7% was achieved, as shown in Table 2.

Table 2 Neural Net Model Training and Validation Result

	Training Set (number of trips)				**Training Set (% of trips)**				**Training Set Performance**	
	Auto	Transit	Active	Total	Auto	Transit	Active	Total	Recall	Precision
Auto	5910	347	539	6796	30.3	1.8	2.8	34.9	87.0%	91.8%
Transit	208	6006	453	6667	1.1	30.8	2.3	34.2	90.1%	85.2%
Active	323	693	5020	6036	1.7	3.6	25.7	31.0	83.2%	83.5%
Total	6441	7046	6012	19499	33.0	36.1	30.8	100	86.9%	86.9%
	Test Set (number of trips)				**Test Set (% of trips)**				**Test Set Performance**	
	Auto	Transit	Active	Total	Auto	Transit	Active	Total	Recall	Precision
Auto	308	35	59	402	14.6	1.7	2.8	19.1	76.7%	77.8%
Transit	58	1206	115	1379	2.8	57.3	5.4	65.5	87.5%	93.0%
Active	30	56	239	325	1.4	2.7	11.3	15.4	73.5%	57.9%
Total	396	1297	412	2106	18.8	61.6	19.6	100	83.3%	83.3 %

The mode probabilities of cellphone traces were then imputed using the trained neural network model. While for all the trips from cellular traces were assigned mode probabilities, in this section, we focus only on trips made during weekdays (to be comparable with MHMS survey data). The overall aggregated results indicate mode shares of 27%, 43%, and 31% for auto, transit and active modes for weekdays. Since there is no ground truth to assess the accuracy of the imputed modes, we use several measures to evaluate the results, by examining the results at different dimensions, temporally and spatially.

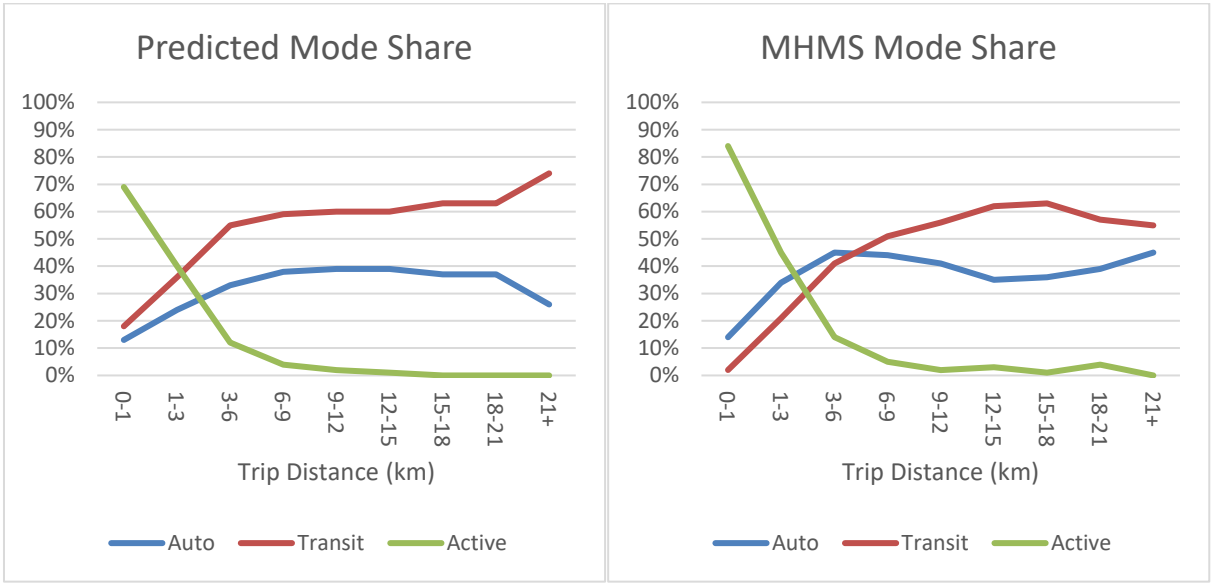

Figure 3. Predicted and Observed Mode Shares by Trip Distance

Figure 3 shows predicted cellular network and observed MHMS trips by mode and distance. The overall trend of the predicted mode shares by the trip distance is similar to the observed one from MHMS interzonal trips. One can note in the figure that short-distance (0-6km) trips are over-predicted for transit and under-predicted for auto travel relative to MHMS data. This confusion is not a particularly surprising result, given the size of the traffic zones used in the analysis and the very simple set of explanatory factors being used in the current model. The results for active mode are, on the other hand, quite good.

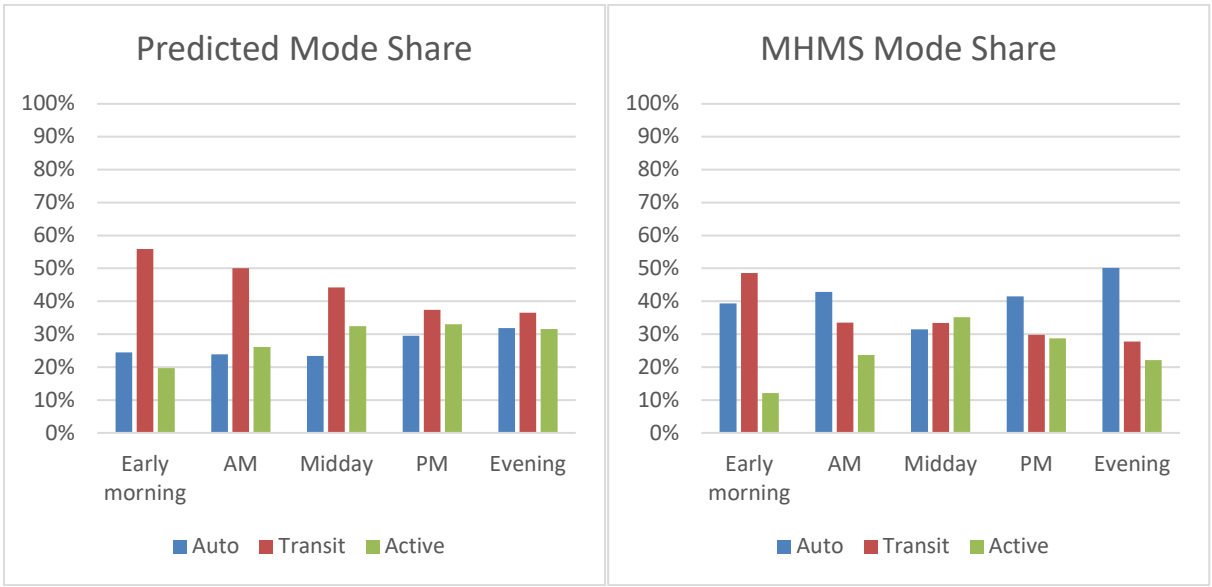

Figure 4. Predicted and Observed Mode Shares by Time of Day

We further examine the predicted modes by time of the day. Figure 4 compares cellphone and MHMS mode shares by time of day (trip start time). Periods considered are Early morning (00:00-05:59), AM peak (06:00-08:59), Mid-day (09:00-14:59), PM peak (15:00-18:59), Evening (19:00-24:00). The patterns are generally similar, although the MHMS data, due to its smaller sample, has a more "fluctuating" trend, especially for early morning and evening hours. Nevertheless, the neural network model produces reasonably similar results.

Figure 5 maps the spatial distribution within the Montevideo region of mode shares for auto, transit and active travel modes, as predicted by the neural network model for the cellphone traces. In all cases, the spatial distributions are plausible, with fewer auto trips and more transit and active trips occurring in more central locations. For transit, there are high mode shares for a few rural areas, which is most likely due to a confusion between transit and auto due to underreporting of trips in those areas. Except for those zones, the overall patterns are reasonable and expected.

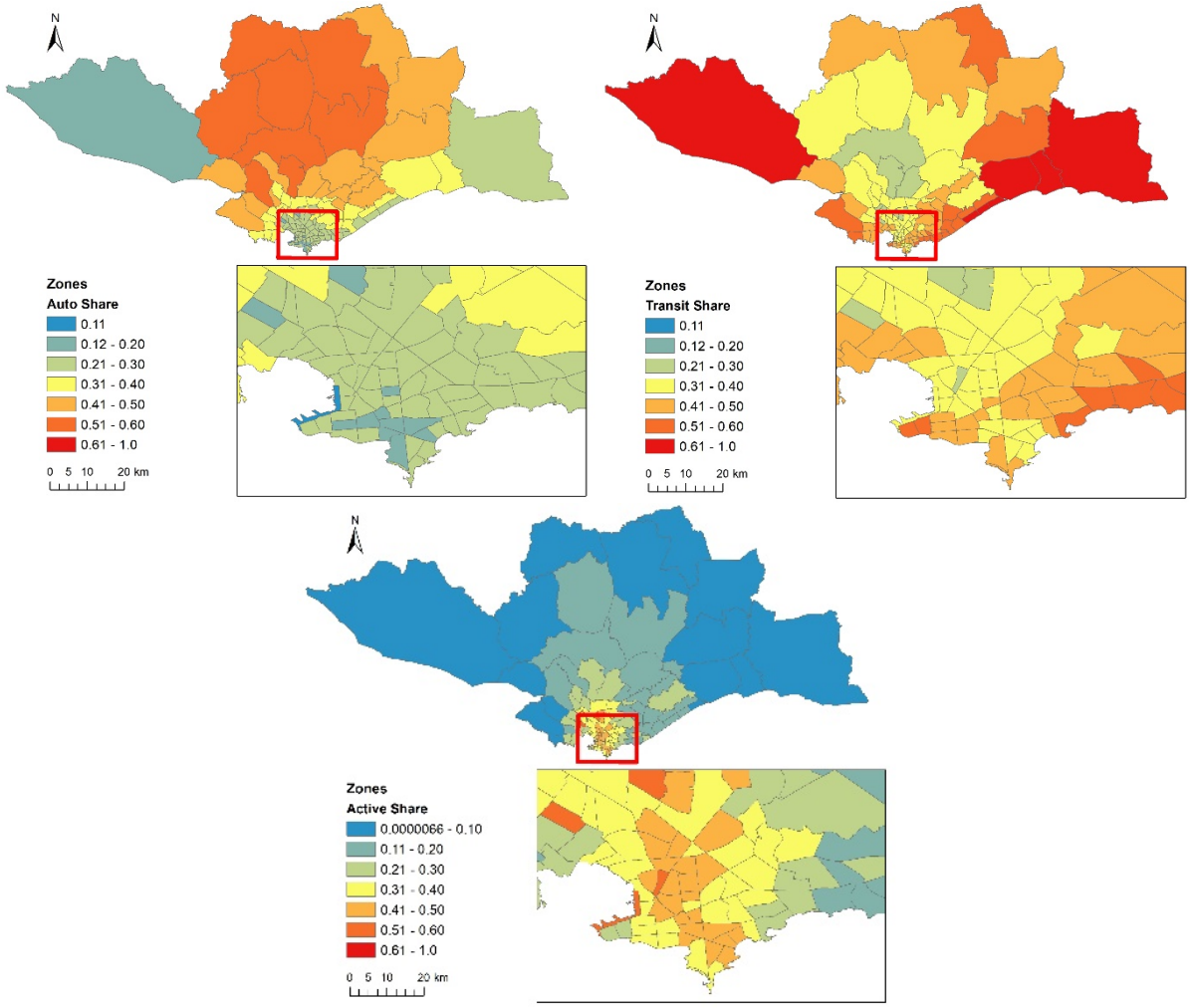

Figure 5. Imputed Mode Share, weekday trips by origin zone

6. Conclusions

This study has developed and presented a novel procedure for fusing cellphone trace data with traditional home interview survey records and public transit smartcard transaction data to infer mode of travel for trips identified from cellphone traces. The results can be used to create an enhanced representation of origin-destination trip-making by time of day and travel mode (Faghih-Imani et al. 2018). To our knowledge this is the first attempt to impute travel modes at the trip level on a large and complete geolocation dataset from a cellular network.

The procedure developed is demonstrated using May, 2018 cellphone and transit smartcard data, along with 2016 MHMS data for Montevideo, Uruguay. It can be used to generate similar datasets for additional periods of time providing that similar cellphone and transit data are available for these time periods. Thus, a time-series of detailed cross-sectional "snapshots" of travel behaviour can be constructed over time (e.g., perhaps on an annual basis). Note that this depends on the assumption that the trip purpose and mode relationships established based on the 2016 MHMS data hold into future time periods. As one moves further into the future, this assumption, of course, will become somewhat more difficult to maintain. Thus, a need for at least an occasional small sample update of survey data over time that provides information concerning trip purposes and modes may well continue.

The study is not without limitations. First, the MHMS data sample size was very small which may cause some bias in our results. Future efforts using a larger dataset to be used as a training set would be recommended to ensure a more accurate representation of mode choices in the region. Second, we assumed a 30-min stationary duration to identify activities. More advanced algorithms (for example, those developed in (Widhalm et al. 2015; Zahedi and Shafahi 2017)) can be used to more confidently detect activities and thus reduce errors in generated trips. Nevertheless, such errors have minimum impact on the proposed algorithmic pipeline for mode imputation. Third, walk and bicycle modes were combined into one "active" mode. Future analysis can further examine these two modes separately. Fourth, alternative machine learning methods to deep neural nets should be explored for possible improvement in performance and/or interpretability. Fifth, an expanded set of explanatory variables ("features") should be tested to determine if they can improve model performance, especially with respect to active mode imputation. Finally, a major limitation of using cellular network data for any transportation analysis is the lack of spatial accuracy, with smaller zones in dense urban areas and larger zones in rural areas. This spatial inaccuracy might generate biases in our analysis, notably possible underestimation of overall trip rates due inability to identify intrazonal trips. In future work, using sensitivity analysis we will explore the effect of zone size on our proposed method. Overall, the presented algorithmic pipeline works reasonably well, and we expect to obtain better results a larger household travel survey becomes available. The procedure presented in this paper also, obviously, is readily extendable to any other urban region possessing similar datasets.

Acknowledgments

This study was funded by the Development Bank of Latin America (CAF). The unstinting support and patience of CAF, and, in particular, Nicolas Estupiñan and Andres Alcala has been greatly appreciated. The collaboration with Diego Hernandez, Universidad Católica del Uruguay, and Antonio Mauttone, Universidad de la República de Uruguay, has been most welcome and helpful. The generous support of Antel in providing access to the cellular data records and of the Intendencia de Montevideo for access to the transit transaction data analyzed in the report is gratefully acknowledged. The assistance of Juan Pablo Pignataro and Juan Andrés Negreira Marin from Antel and Verónica Orellano Chiazzaro and Leonardo Goday from the Intendencia has been much appreciated. The opinions expressed in this paper are those of the authors and do not necessarily reflect those of CAF, Antel or the Indendencia de Montevideo.

References

Ahas, R., Silm, S., Järv, O., Saluveer, E., Tiru, M.: Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. J. Urban Technol. J. ISSN homep, 1063–732 (2010). https://doi.org/10.1080/10630731003597306

Alexander, L., Jiang, S., Murga, M., González, M.C.: Origin–destination trips by purpose and time of day inferred from mobile phone data. Transp. Res. Part C Emerg. Technol. 58, 240–250 (2015). https://doi.org/10.1016/J.TRC.2015.02.018

Axhausen, K.W.: Learning from smart cards: Lessons from Singapore. (2017). https://doi.org/10.3929/ethz-b-000197229

Bachir, D., Khodabandelou, G., Gauthier, V., El Yacoubi, M., Puchinger, J.: Inferring dynamic origin-destination flows by transport mode using mobile phone data. Transp. Res. Part C Emerg. Technol. 101, 254–275 (2019). https://doi.org/10.1016/J.TRC.2019.02.013

Bar-Gera, H.: Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel. Transp. Res. Part C Emerg. Technol. 15, 380–391 (2007). https://doi.org/10.1016/J.TRC.2007.06.003

Becker, R.A., Caceres, R., Hanson, K., Loh, J.M., Urbanek, S., Varshavsky, A., Volinsky, C.: A Tale of One City: Using Cellular Network Data for Urban Planning. IEEE Pervasive Comput. 10, 18–26 (2011). https://doi.org/10.1109/MPRV.2011.44

Bekhor, S., Cohen, Y., Solomon, C.: Evaluating long-distance travel patterns in Israel by tracking cellular phone positions. J. Adv. Transp. 47, 435–446 (2013). https://doi.org/10.1002/atr.170

Bekhor, S., Shem-Tov, I.B.: Investigation of travel patterns using passive cellular phone data. J. Locat. Based Serv. 9, 93–112 (2015). https://doi.org/10.1080/17489725.2015.1066515

Berlingerio, M., Calabrese, F., Di Lorenzo, G., Nair, R., Pinelli, F., Sbodio, M.L.: AllAboard: A System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data. Presented at the (2013)

Bohte, W., Maat, K.: Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. Transp. Res. Part C Emerg. Technol. 17, 285–297 (2009). https://doi.org/10.1016/J.TRC.2008.11.004

Bonnel, P., Hombourger, E., Olteanu-Raimond, A.-M., Smoreda, Z.: Passive Mobile Phone Dataset to Construct Origin-destination Matrix: Potentials and Limitations. Transp. Res. Procedia. 11, 381–398 (2015). https://doi.org/10.1016/J.TRPRO.2015.12.032

Broach, J., Dill, J., McNeil, N.W.: Travel mode imputation using GPS and accelerometer data from a multi-day travel survey. J. Transp. Geogr. 78, 194–204 (2019). https://doi.org/10.1016/J.JTRANGEO.2019.06.001

Caceres, N., Romero, L.M., Benitez, F.G.: Inferring origin-destination trip matrices from aggregate volumes on groups of links: a case study using volumes inferred from mobile phone data. J. Adv. Transp. 47, 650–666 (2013). https://doi.org/10.1002/atr.187

Caceres, N., Romero, L.M., Benitez, F.G., del Castillo, J.M.: Traffic Flow Estimation Models Using Cellular Phone Data. IEEE Trans. Intell. Transp. Syst. 13, 1430–1441 (2012). https://doi.org/10.1109/TITS.2012.2189006

Caceres, N., Wideberg, J.P., Benitez, F.G.: Deriving origin–destination data from a mobile phone network. IET Intell. Transp. Syst. 1, 15 (2007). https://doi.org/10.1049/iet-its:20060020

Caceres, N., Wideberg, J.P., Benitez, F.G.: Review of traffic data estimations extracted from cellular networks. IET Intell. Transp. Syst. 2, 179 (2008). https://doi.org/10.1049/iet-its:20080003

Calabrese, F., Di Lorenzo, G., Liu, L., Ratti, C.: Estimating Origin-Destination Flows Using Mobile Phone

Location Data. IEEE Pervasive Comput. 10, 36–44 (2011). https://doi.org/10.1109/MPRV.2011.41

Chapleau, R., Gaudette, P., Spurr, T.: Strict and Deep Comparison of Revealed Transit Trip Structure between Computer-Assisted Telephone Interview Household Travel Survey and Smart Cards. Transp. Res. Rec. J. Transp. Res. Board. 2672, 13–22 (2018). https://doi.org/10.1177/0361198118758297

Chen, C., Bian, L., Ma, J.: From traces to trajectories: How well can we guess activity locations from mobile phone traces? Transp. Res. Part C Emerg. Technol. 46, 326–337 (2014). https://doi.org/10.1016/j.trc.2014.07.001

Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M.: The promises of big data and small data for travel behavior (aka human mobility) analysis, (2016)

Çolak, S., Alexander, L.P., Alvim, B.G., Mehndiratta, S.R., González, M.C.: Analyzing Cell Phone Location Data for Urban Travel. Transp. Res. Rec. J. Transp. Res. Board. 2526, 126–135 (2015). https://doi.org/10.3141/2526-14

Demissie, M.G., Antunes, F., Bento, C., Phithakkitnukoon, S., Sukhvibul, T.: Inferring origin-destination flows using mobile phone data: A case study of Senegal. In: 2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). pp. 1–6. IEEE (2016)

Doyle, J., Hung, P., Kelly, D., Mcloone, S., Farrell, R.: Utilising Mobile Phone Billing Records for Travel Mode Discovery. In: ISSC 2011 (2011)

Faghih-Imani, A., Vaughan, J., Yusuf, B., Miller, E.J.: Final Project Report, Report 6: iCity-South: Urban Informatics for Sustainable Metropolitan Growth in Latin America. (2018)

Graells-Garrido, E., Caro, D., Parra, D.: Inferring modes of transportation using mobile phone data. EPJ Data Sci. 7, 49 (2018). https://doi.org/10.1140/epjds/s13688-018-0177-1

Gundlegård, D., Rydergren, C., Breyer, N., Rajna, B.: Travel demand estimation and network assignment based on cellular network data. Comput. Commun. 95, 29–42 (2016). https://doi.org/10.1016/J.COMCOM.2016.04.015

Hillel, T., Bierlaire, M., Jin, Y.: A systematic review of machine learning methodologies for modelling passenger mode choice. (2019)

Huang, H., Cheng, Y., Weibel, R.: Transport mode detection based on mobile phone network data: A systematic review. Transp. Res. Part C Emerg. Technol. 101, 297–312 (2019). https://doi.org/10.1016/J.TRC.2019.02.008

Hui, K.T.Y., Wang, C., Kim, A., Qiu, T.Z.: Investigating the Use of Anonymous Cellular Phone Data to Determine Intercity Travel Volumes and Modes. In: Transportation Research Board 96th Annual Meeting (2017)

Iqbal, M.S., Choudhury, C.F., Wang, P., González, M.C.: Development of origin–destination matrices using mobile phone call data. Transp. Res. Part C Emerg. Technol. 40, 63–74 (2014). https://doi.org/10.1016/J.TRC.2014.01.002

Janecek, A., Valerio, D., Hummel, K.A., Ricciato, F., Hlavacs, H.: The Cellular Network as a Sensor: From Mobile Phone Data to Real-Time Road Traffic Monitoring. IEEE Trans. Intell. Transp. Syst. 16, 2551–2572 (2015). https://doi.org/10.1109/TITS.2015.2413215

Janzen, M., Axhausen, K.W.: Long-Term-C-TAP Simulation: Generating Long Distance Travel Demand for a full Year. (2015)

Janzen, M., Vanhoof, M., Smoreda, Z., Axhausen, K.W.: Closer to the total? Long-distance travel of French mobile phone users. Travel Behav. Soc. 11, 31–42 (2018). https://doi.org/10.1016/j.tbs.2017.12.001

Järv, O., Ahas, R., Saluveer, E., Derudder, B., Witlox, F.: Mobile Phones in a Traffic Flow: A Geographical Perspective to Evening Rush Hour Traffic Analysis Using Call Detail Records. PLoS One. 7, e49171

(2012). https://doi.org/10.1371/journal.pone.0049171

Jiang, S., Fiore, G.A., Yang, Y., Ferreira, J., Frazzoli, E., González, M.C.: A review of urban computing for mobile phone traces. In: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing - UrbComp '13. p. 1. ACM Press, New York, New York, USA (2013)

Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. (2014)

Koushik, A.N., Manoj, M., Nezamuddin, N.: Machine learning applications in activity-travel behaviour research: a review. Transp. Rev. 40, 288–311 (2020). https://doi.org/10.1080/01441647.2019.1704307

Larijani, A.N., Olteanu-Raimond, A.-M., Perret, J., Brédif, M., Ziemlicki, C.: Investigating the Mobile Phone Data to Estimate the Origin Destination Flow and Analysis; Case Study: Paris Region. Transp. Res. Procedia. 6, 64–78 (2015). https://doi.org/10.1016/J.TRPRO.2015.03.006

Lee, R.J., Sener, I.N., Mullins, J.A.: An evaluation of emerging data collection technologies for travel demand modeling: from research to practice. Transp. Lett. 8, 181–193 (2016). https://doi.org/10.1080/19427867.2015.1106787

Mellegard, E., Moritz, S., Zahoor, M.: Origin/Destination-estimation Using Cellular Network Data. In: 2011 IEEE 11th International Conference on Data Mining Workshops. pp. 891–896. IEEE (2011)

Miller, E.J., Parada Hernandez, C., Habib, K.N.: Review of the Montevideo Home Mobility Survey, Report 2: iCity-South: Urban Informatics for Sustainable Metropolitan Growth in Latin America. (2017)

Nanni, M., Trasarti, R., Furletti, B., Gabrielli, L., Mede, P. Van Der, Bruijn, J. De, Romph, E. De, Bruil, G.: Transportation Planning Based on GSM Traces: A Case Study on Ivory Coast. Presented at the (2014)

Ordóñez Medina, S.A.: Inferring weekly primary activity patterns using public transport smart card data and a household travel survey. Travel Behav. Soc. 12, 93–101 (2018). https://doi.org/10.1016/j.tbs.2016.11.005

Pelletier, M.P., Trépanier, M., Morency, C.: Smart card data use in public transit: A literature review. Transp. Res. Part C Emerg. Technol. 19, 557–568 (2011). https://doi.org/10.1016/j.trc.2010.12.003

Phithakkitnukoon, S., Sukhvibul, T., Demissie, M., Smoreda, Z., Natwichai, J., Bento, C.: Inferring social influence in transport mode choice using mobile phone data. EPJ Data Sci. 6, 11 (2017). https://doi.org/10.1140/epjds/s13688-017-0108-6

Pozdnoukhov, A., Campbell, A., Feygin, S., Yin, M., Mohanty, S.: San Francisco Bay Area: The SmartBay Project - Connected Mobility. In: The Multi-Agent Transport Simulation MATSim. pp. 485–490. Ubiquity Press (2016)

Pucci, P., Manfredini, F., Tagliolato, P.: Daily Mobility Practices Through Mobile Phone Data: An Application in Lombardy Region. Presented at the (2015)

Qu, Y., Gong, H., Wang, P.: Transportation Mode Split with Mobile Phone Data. In: 2015 IEEE 18th International Conference on Intelligent Transportation Systems. pp. 285–289. IEEE (2015)

Schlaich, J., Otterstätter, T., Friedrich, M.: Generating Trajectories from Mobile Phone Data. In: Transportation Research Board 89th Annual MeetingTransportation Research Board (2010) (2010)

Semanjski, I., Gautama, S., Ahas, R., Witlox, F.: Spatial context mining approach for transport mode recognition from mobile sensed big data. Comput. Environ. Urban Syst. 66, 38–52 (2017). https://doi.org/10.1016/J.COMPENVURBSYS.2017.07.004

Smoreda, Z., Olteanu-Raimond, A.-M., Couronné, T.: Spatiotemporal Data from Mobile Phones for Personal Mobility Assessment. In: Transport Survey Methods: Best Practice for Decision Making. pp. 745–768. Emerald Group Publishing Limited (2013)

Sohn, T., Varshavsky, A., LaMarca, A., Chen, M.Y., Choudhury, T., Smith, I., Consolvo, S., Hightower, J., Griswold, W.G., de Lara, E.: Mobility Detection Using Everyday GSM Traces. Presented at the (2006)

Tettamanti, T., Varga, I.: Mobile Phone Location Area Based Traffic Flow Estimation in Urban Road

Traffic. Columbia Int. Publ. Adv. Civ. Environ. Eng. 1, 1–15 (2014)

Toole, J.L., Colak, S., Sturt, B., Alexander, L.P., Evsukoff, A., González, M.C.: The path most traveled: Travel demand estimation using big data resources. Transp. Res. Part C Emerg. Technol. 58, 162–177 (2015). https://doi.org/10.1016/J.TRC.2015.04.022

Wang, H., Calabrese, F., Di Lorenzo, G., Ratti, C.: Transportation mode inference from anonymized and aggregated mobile phone call detail records. In: 13th International IEEE Conference on Intelligent Transportation Systems. pp. 318–323. IEEE (2010)

Wang, Z., He, S.Y., Leung, Y.: Applying mobile phone data to travel behaviour research: A literature review. Travel Behav. Soc. 11, 141–155 (2018). https://doi.org/10.1016/J.TBS.2017.02.005

Widhalm, P., Yang, Y., Ulm, M., Athavale, S., González, M.C.: Discovering urban activity patterns in cell phone data. Transportation (Amst). 42, 597–623 (2015). https://doi.org/10.1007/s11116-015-9598-x

Wu, C., Thai, J., Yadlowsky, S., Pozdnoukhov, A., Bayen, A.: Cellpath: Fusion of cellular and traffic sensor data for route flow estimation via convex optimization. Transp. Res. Part C Emerg. Technol. 59, 111–128 (2015). https://doi.org/10.1016/J.TRC.2015.05.004

Xu, D., Song, G., Gao, P., Cao, R., Nie, X., Xie, K.: Transportation Modes Identification from Mobile Phone Data Using Probabilistic Models. Presented at the December 17 (2011)

Xu, Y., Shaw, S.-L., Zhao, Z., Yin, L., Fang, Z., Li, Q.: Understanding aggregate human mobility patterns using passive mobile phone location data: a home-based approach. Transportation (Amst). 42, 625–646 (2015). https://doi.org/10.1007/s11116-015-9597-y

Xu, Y., Shaw, S.-L., Zhao, Z., Yin, L., Lu, F., Chen, J., Fang, Z., Li, Q.: Another Tale of Two Cities: Understanding Human Activity Space Using Actively Tracked Cellphone Location Data. Ann. Am. Assoc. Geogr. 106, (2017). https://doi.org/10.1080/00045608.2015.1120147

Yin, M., Sheehan, M., Feygin, S., Paiement, J.-F., Pozdnoukhov, A.: A Generative Model of Urban Activities from Cellular Data. IEEE Trans. Intell. Transp. Syst. 1–15 (2017). https://doi.org/10.1109/TITS.2017.2695438

Zahedi, S., Shafahi, Y.: Estimating activity patterns using spatio-temporal data of cell phone networks. Int. J. Urban Sci. 1–18 (2017). https://doi.org/10.1080/12265934.2017.1331139

Zhang, Y., Martens, K., Long, Y.: Revealing group travel behavior patterns with public transit smart card data. Travel Behav. Soc. 10, 42–52 (2018). https://doi.org/10.1016/j.tbs.2017.10.001

Zhang, Y., Qin, X., Dong, S., Ran, B.: Daily O-D Matrix Estimation Using Cellular Probe Data. Presented at the (2010)