Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2019) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

22nd EURO Working Group on Transportation Meeting, EWGT 2019, 18-20 September 2019, Barcelona, Spain

Using GPS tracking data to validate route choice in OD trips within dense urban networks

Lídia Montero[a *] and Xavier Ros-Roca[a,b]

[a]Universitat Politècnica de Catalunya (UPC) Carrer Jordi Girona 1-3 08034 Barcelona Spain
[b]PTV Iberia Plaça Catalunya 1 08002 Barcelona Spain

Abstract

There are presently several companies that provide processed or raw global positioning system (GPS) measurements generated by fleets of commercial vehicles, internet applications or automobile companies. The aim of this paper is to deepen our understanding of GPS data applicability in transportation modelling by providing systematic and quantified insights into the representativeness of collected data in describing individual' route choices. Unfortunately, real data often contain noise, uncertainty, errors, redundancies or even irrelevant information. Useless models will be obtained when built over incorrect or incomplete data. This is why pre-processing is one of the most critical steps in data analysis. Yet, pre-processing has not been properly systematized, which is why this paper focuses on the necessary steps for pre-processing GPS tracking data together and puts forth a proposal for systematizing it. Furthermore, the aggregation level is at waypoint location for low latency GPS positions along the trip trajectory. Travel time reliability on OD paths is addressed, along with other OD path characteristics. Dense urban networks pose multiple possibilities for route choice, and data from new technologies offers an opportunity to understand route choice behaviour.

Peer-review under responsibility of the scientific committee of the 22nd EURO Working Group on Transportation Meeting.
Keywords: Type your keywords here, separated by semicolons ;

1. Introduction

INRIX (http://inrix.com/) gathers real-time, predictive and historical data from different sources of particular low-latency GPS tracking data (Cerqueira et al. 2018). In this way, it facilitates the automatic extraction of valuable

* Lídia Montero. Tel.: +34-93-4011038; fax: +34-93-4015855.
E-mail address: lidia.montero@upc.edu

Peer-review under responsibility of the scientific committee of the 22nd EURO Working Group on Transportation Meeting.

mobility information through distinct data mining processes. In the past, the authors (Montero et al. 2019) have been involved in the estimation of modal dynamic OD matrices obtained from CDRs (Call Detail Records) compiled from Orange customers (Toole et al. 2015). For decades, transportation researchers have largely used data from active data requests, such as travel surveys in which subjects self-report their activities and travel via paper, web, or phone interviews; travel surveys coupled with GPS loggers, in which subjects both complete questionnaires and carry GPS loggers; and pure GPS-based surveys, in which subjects only carry GPS loggers (Chen et al. 2016; Montini, Antoniou, and Axhausen 2017; Shen and Stopher 2014). In this work, we focus on passive data, meaning data not collected through active data requests; rather, it is generated for purposes that are not intended for research but can potentially be used for it. The primary objective of this paper is to discuss GPS tracking data that is passively collected from private and fleet vehicles using INRIX traffic information, which at the same time produces GPS data that is delivered to the company's analytics processes. In particular, INRIX GPS tracking data for the first crown of the Barcelona metropolitan area is comprises roughly 50% of passive data from consumer vehicles. The UPC is developing and calibrating a transportation planning model in VISUM as a part of the Virtual Mobility Lab (VML), a strategic project aiming to address and evaluate new scenarios involving MaaS (Mobility as a Service) services that are funded by CARNET (http://www.carnetbarcelona.com/).

In transportation research, discrete choice models (Ben-Akiva and Lerman 1985) are developed to predict travel behaviours, that is, more specifically choices on modes of transportation, destinations, and routes. Discrete choice models are developed based on the theory of random utility maximization (RUM). RUM assumes that when choosing from a set of discrete alternatives (e.g., mode choices and destination choices), the decision maker is aware of all feasible alternatives and their associated attributes, is willing to make trade-offs across attributes and, given these conditions, he or she then chooses the choice that will maximize his or her satisfaction. The standard discrete choice models (e.g., a Multinomial Logit Model) usually hold some common assumptions, i.e.: (1) the random components of the utilities of the different alternatives are independent and identically distributed (IID) with a type I extreme-value (or Gumbel) distribution; (2) homogeneous responsiveness to attributes of alternatives is maintained across individuals (i.e., an assumption of response homogeneity); and (3) the error variance–covariance structure of the alternatives is identical across individuals (i.e., an assumption of error variance–covariance homogeneity) (Hensher and Greene 2003). However, these assumptions are easily violated in practice, usually when a variable that has not been considered in the systematic part of the model or an unobservable variable affects route choice, as for example the opportunity of car-sharing with a colleague. The discovery of certain mobility patterns from GPS tracking offers an opportunity to identify the links between microscopic individual choices and macroscopic behaviours observed from traffic counts and to re-examine the decision rules used to model OD path-related choices. Despite the widespread use of RUM models, some experiments in behavioural studies have found deviations from their axioms, and this has led to the development of Non-Expected Utility Theories (de Moraes-Ramos, Daamen, and Hoogendoorn 2013).

Prior to our path use study, we plan to address the stability of OD path choice and level of congestion. OD paths will be identified and characterized in terms of time, distance and vehicle class.

2. Experimental Study Area

The selected site for the computational experiments is the first crown of the Barcelona metropolitan area. It is composed of 18 municipalities with 2,837,000 inhabitants, the highest population concentration (around 60%) in the metropolitan area. Its primary road and public transportation network contains more than 200 bus lines with over 4,000 stops, 10 metro lines, 15 railways lines, and 2 tramway networks.

GPS devices can record the travel time and the coordinates of locations with low latency, which can therefore report speed, start time, end time, and the routes of trips. However, GPS devices cannot automatically identify trip ends or report travel modes and trip purposes. INRIX sources of GPS data are fleets of commercial vehicles and private cars. Therefore, travel mode is known by default, but trip purposes for private consumers are not available and identification of trip end is somewhat arbitrary, since it uses some common established rules (10 min of inactivity or motion within a 100 m radius). Trip concept does not apply to fleets, and delivery circuits are useful only for monitoring traffic conditions. But they are not suitable for either addressing the origin-destination (OD) spatial distribution of trips or validating OD path route choice models.

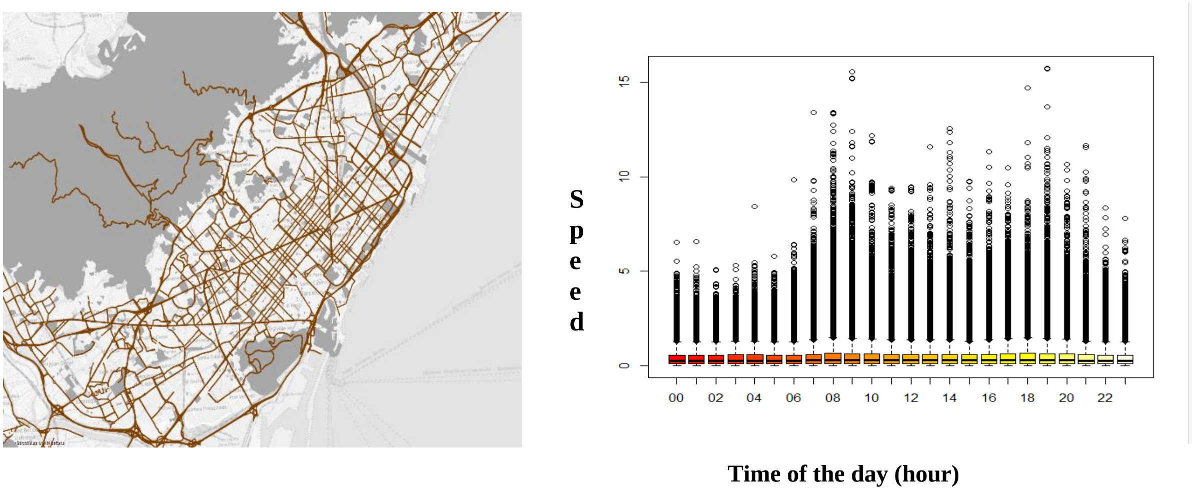

Fig. 1. (a) First crown of Barcelona metropolitan area; (b) INRIX XD Segment hourly speeds (km/h) for internal trips

One month (March 2017) of INRIX GPS tracking datasets were used. The raw data consists of 846,295 trips (internal for origin/destination in the study area) and 61,925,267 waypoints in .csv files of, respectively, 236 MB (247,534,055 bytes) and 1.78 GB (1,912,468,534 bytes). Additionally, speed profiles for INRIX XD Segments (7,964 for the study area, see Fig. 1. (a)) are available at a temporal resolution of 1 min and travel times can be derived. The average length of an XD Segment is about 200m, with 1 km being the longest, and it usually follows main arteries or major streets in the traffic network. The target population addressed in this study are representative drivers of internal trips in the Barcelona metropolitan area. Raw data from flat .csv files had to be filtered, transformed and aggregated/grouped before the actual data analysis could take place. Once the data have been restructured using the most basic methods of data science and analytics, there are several other more complex tasks such as splitting, applying and joining data, computing table margins, and casting/merging data. Despite the sheer vastness of current research and practice on big data, its role in transportation remains underexplored, particularly in terms of needs and available opportunities (Antoniou, Dimitriou, and Pereira 2019) (Chodrow et al. 2016). Real data sets tend to be imperfect, contain errors, outliers, missing data, and extra noise. Tools are required for either detecting or correcting them, as the application of certain analytics techniques may require specific conditions for the data set (only binary variables, centred data, normality, only qualitative variables, etc.). In this case, they would be tools for verifying that those conditions hold, or eventually transforming data appropriately to meet those conditions. Data preparation tasks are often time consuming and difficult, and few papers in the literature address this critical topic; mainly because the approaches taken should be tailored to each specific application and human interaction is required. The authors are conscious of the importance of very careful and rigorous pre-processing, for which we have dedicated sufficient time to these efforts.

A clear and detailed methodology for tackling pre-processing is described in (Gibert, Sànchez–Marrè, and Izquierdo 2016). Pre-processing is supported by intensive use of basic descriptive statistics and basic univariate and bivariate graphical representations of data. Some multivariate descriptive techniques (Principal Component Analysis) are also useful detecting high-order outliers or some nonlinearities. Some more or less sophisticated visualization techniques play a crucial role. Among all suggested pre-processing techniques, the most critical are detection techniques oriented toward detecting imperfections in datasets or verifying that required assumptions are fulfilled in a particular analysis. These are often associated with some kind of diagnoses on data and, consequently, to some decisions related to pre-processing: outlier detection, data error detection, missing data detection, detecting relevance or redundancy, feature weighting, independence assessment, detection of influential observations, and normality/linearity assessment. Transforming techniques transform the dataset in order to correct the previously detected imperfections or fulfil the technical conditions necessary for certain analytical techniques, such as: determining the active set of data matrix rows and columns (by selecting expert-based objects and variables, and filtering), outlier treatment, error data treatment, missing data treatment, treating relevance and redundancy, dimensionality reduction techniques, instance selection, resampling, feature selection, factorial methods, and transforming variables (through homogenization, differences and ratios, compositional data, and functional transformations that are inverse, logarithmic, etc.). Recodification usually involves deriving new variables, and this is achieved by discretization, centring, standardization, normalization and, finally, the creation of new variables through aggregation, feature extraction, building indicators and dealing with multivalued variables.

3. Methodology

Pre-processing methodological steps were applied to INRIX data and meta-data provided by INRIX was used to define proper variable types (quantitative, qualitative, time variables). Trip attributes are shown in Table 1. The distribution of provider type depends on the geospatial type. Focusing on internal trips to Barcelona's Metropolitan area, almost 53.9% of the total trips on weekdays pertain to passenger cars (Audi and other expensive cars). Almost 70% of the trips pertain to EndpointType 0, indicating that the trip does not start or end at a stop; therefore the trip does not start/end at any origin origin/destination and thus the trip trajectory does not correspond to an origin-destination trip but is instead a trip leg. **Therefore, the sample of OD trips included in INRIX does not correspond to the OD trip pattern for internal trips in the study area.** Trips are composed of waypoints, GPS coordinates (given by longitude and latitude coordinates, as well as several important additional fields). Latency depends on the provider, but it is usually less than 10 sec in Barcelona's Metropolitan data – according to INRIX. Fields included for each waypoint register are: trip identifier, waypoint sequence, date, time, longitude, latitude, segment identifier, ZoneName, DeviceId, RawSpeed, RawSpeedMetric and LinkID. Building the working data matrix consists of determining the target population defined for the analysis, for which the data registers (instances) and attributes of this population will be kept. *Filtering* is devoted mainly to selecting subsamples from the main data matrix in order to restrict the scope of the analysis and eliminate observations from other domains that are not targeted. Since different behavior and traffic conditions are commonly assumed, the target population was selected according to the following filters:

- A subset of trip registers pertaining to working days (728,060 out of 846,295 trips).
- A subset of internal trips to the first crown of Barcelona's metropolitan area (456,751 out of 728,060 trips).
- A subset of trips pertaining to private INRIX data consumers (245,728 out of 456,751 trips).
- Waypoint data for working trips (16,065,344 out 61,925,267 waypoints).
- Removed attributes of trip data: Mode, IsStartHome, IsEndHome, ProbeSourceType, MultipleCorridors, MultipleZones, MovementType, OriginCbg, DestCbg, GeospatialType, ProviderType, ProviderDrivingProfile and VehicleweightClass. 16 columns were retained.
- Removed attributes of waypoint data: FRC, RawSpeedMetric. 10 attributes retained.
- Data types and levels for factors were recoded with meaningful labels for trip and waypoint data matrices.
- Trip travel time in minutes was calculated for working trips and stored in a new column of trip data matrix.

All the initially preserved columns were summarized using RStudio. The first summary indicated the missing values appearing for trip data under OriginZoneName, DestinationZoneName and for waypoint data under ZoneName, SegmentId and LinkId. After selecting the working data, some attributes of trip data became non-informative with constant values: GeospatialType, ProviderType, ProviderDrivingProfile, vehicleweightClass and MovementType. Related columns and RawSpeedMetric for the waypoint data matrix (set to km/h) were removed.

Table 1. INRIX trip attributes. Only available fields for Barcelona's data.

TripID	A trip's unique identifier (string)
DeviceID	A device's unique identifier (string)
ProviderID	A provider's unique identifier (string)
Mode	[alpha version -- do not use]
StartDate	The trip's start date and time in UTC, ISO-8601 format, example: "2014-04-01T08:33:35.000Z" (string)
StartWDay	[deprecated] Values from 1 to 6 (Monday to Sunday) (integer)
EndDate	The trip's end date and time in UTC, ISO-8601 format, example: "2014-04-01T08:33:35.000Z" (string)
EndWDay	[deprecated] Values from 1 to 6 (Monday to Sunday) (integer)
StartLocLat	The latitude coordinates of the trip's start point in decimal degrees (floating point f6.3)
StartLocLon	The longitude coordinates of the trip's start point in decimal degrees (floating point f6.3)
EndLocLat	The latitude coordinates of the trip's end point in decimal degrees (floating point f6.3)
EndLocLon	The decimal degree longitude coordinates of the trip's end point in decimal degree (floating point f6.3)
GeospatialType	Describes the trip's geospatial intersection with the requested zones. Polythomic factor with 4 levels (EE, EI, IE, II)
ProviderType	Numeral representing the provider type (Consumer, Fleet, Mobile) – Polythomic factor
ProviderDrivingProfile	Numeral representing the provider driving profile. Polythomic factor (Consumer, Taxi, LocalDeliver and Trucks)
VehicleWeightClass	Numeral representing the vehicle weight class. Polythomic factor (Cars, Vans and Heavy Trucks)
OriginZoneName	The origin zone of the trip, if the trip started in a zone – this must be filled in (linked to VISUM-VML TAZ – Zones) (integer)
DestinationZoneName	The destination zone of the trip, if the trip ends in a zone – this must be filled in (linked to VISUM-VML TAZ – Zones) (integer)
EndpointType	Indicates if the trip starts and ends at a detected stop (blank=unknown (prior to 2017), -1 = Unknown, 0 = Trip does not start or end at stop, 1 = Trip starts at stop, 2 = Trip ends at stop, 3 = Trip starts and ends at stop). Polythomic factor 5 levels.
TripMeanSpeedKph	Average speed (km/h) – Floating point data (5 decimals)
TripMaxSpeedKph	Max speed (km/h) – Floating point data (5 decimals)
TripDistanceMeters	Trip distance (m) – Floating point data (1 decimal)
MovementType	1 = Moving Trip, 0 = Non-moving Trip – Binary factor
tt.min	Travel time (min) – *Field added by UPC – Floating point (4 decimals)*

OriginZoneName and DestinationZoneName were imputed into the trip working data matrix and ZoneName into the waypoint working data matrix. This was based on projecting onto the Transportation Analysis Zone (TAZ) a shapefile loaded in RStudio with sf package and using coordinate points in the trip and waypoint registers. 27,821 trips were removed due to unsuccessful matching of origin or destination in the TAZ. Intrazonal trips were also discarded, leading to **185,432 trips whose trajectories account for 13,005,532 waypoints**. Imputation of SegmentID and LinkID in waypoint registers was initially supported by the maptools RStudio package and snapPointsToLines() function. XD-Segment shapefile by INRIX and VISUM-VML link shapefile were uploaded. After checking imputation results, network matching was not validated. Then, after several refinements, we continued to work with LinkID imputation because it still presented a lot of inconsistencies. PostGIS (Anon 2018), was used to define map-matching of the nearest XD-Segment to a GPS track register (longitude, latitude, time), but the authors are not fully satisfied with the results.

Meanwhile, in order to prepare the two data matrices for GPS tracking data, we performed outlier detection for trip travel time, trip mean speed and trip distance. We also analyzed the number of waypoints per trip, waypoint latency and number of trips per device. Large latencies make it difficult to identify the paths used in the network when solving the network matching problem. Very short trips in terms of distance and time are not suitable for the goals of this work, as it focuses on OD spatial distribution stability over time and OD route selection. Thus, they were removed from the working data set, as suggested by some authors (Montini et al. 2017). Some statistics about trip and waypoint were set up in the final working sample:

- Trip distance: More than 25% of the trips have a total distance of less than 1km and 5% are greater than 18km. Trip distance median is 2,490m. A right skewed distribution is observed. Only trips with a total distance between 1.5km and 24km are initial candidates for the working set. 70,264 trips are candidates to be either outliers or non-interesting trips for the purpose of the current work.
- Trip Mean Speed (km/h): Median is 16.85 km/h; the 2.5% and 97.5% percentiles are 3.75 and 74.51 km/h, respectively. A very interesting bimodal profile distribution found for mean speed suggests conducting a deeper analysis. Non-supervised clustering is carried out. Trip mean speeds greater than 120km/h (Catalan motorway speed limit) are candidates for outliers (204 trips).
- Trip Travel time (min): Median is 9.61min, which is a really short trip duration; 2.5% and 97.5% percentiles are 3.75 and 74.51 min, respectively. A right-skewed distribution is found. A bivariate boxplot tool is used to discover some insights about atypical trip travel times and trip distances. Only large travel times might cause problems to the goals of this paper, the outlier detection procedure targets as potential outliers those trips whose durations are over 120min (less than 1% of the working sample, 536 trips).
- Number of trips per device and waypoints per trip: Median for number of trips per device is 1; 74% of the devices register just 1 trip and 95% of devices registered 8 trips or less. Latency (sec) between waypoints per trip. Working sample of waypoints per trip might be surprisingly low, percentiles are shown in Table 2 (a) (median is 27 waypoints per trip).
- Waypoint latency: Latency for GPS tracking in the working sample has a median of 5 seconds; percentiles are shown in Table 2 (b) (mean is 10.12 sec). 46,922 trips have a trajectory that includes some latency of greater than 3 min between waypoints, which we assumed to be a threshold for identifying the selected path.

Table 2. (a) Number of waypoints per trip. (b) Latency (sec) between waypoints per trip. Working sample.

	Percentiles – Number of Waypoints per Trip – Working sample										
Prob.	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Number	2	2	8	12	18	27	47	82	133	221	4,914

	Percentiles – Latency between waypoints per Trip (sec) – Working sample										
Prob.	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Seconds	1	1	4	5	5	5	5	5	5	10	3,373

The final working sample for obtaining OD matrices at different spatial aggregations contains 68,246 trips and 6,758,275 waypoints. Only 10,505 trips correspond to OD pairs that have more than 1 trip. Sparsity is common in OD matrices, but if we disregard the TAZ spatial resolution and consider instead a district in municipality resolution, then 73% of OD pairs have more than 1 trip while about 9% have more than 50 trips. Since congestion level has to be controlled, a speed profile analysis was used to classify day-type and hour according to similar congestion. The next step consisted of aggregating trips in OD pairs between districts (in municipalities) pertaining to the same congestion pattern, then selecting those OD pairs containing more trips to address OD path preferences. In the 8 to 10 am period, only 22 OD pairs satisfy the condition of having more than 50 OD trips and they account for 1,086 total trips and 108,498 waypoints.

Once the pre-processing of GPS tracking data is systematized, the aim is to focus on the stability of non-adaptive path choice models between some selected OD pairs containing the largest number of trips. Examples such as Path Size Logit (Ben-Akiva and Bierlaire 1999), C-Logit (Cascetta, Russo, and Vitetta 1997) and (Frejinger, Bierlaire, and Ben-Akiva 2009), among others, have been proposed in literature.

An analysis of common XD-Segments OD paths for the selected OD pairs has been addressed accounting for path size defined as one minus the proportion of common segments, one minus the proportion of common segments weighted by segment length and path size measure defined by (Montini et al. 2017). Path size ranges from 1, for paths that do not overlap with any other in the choice set, to very small values if there is a lot of overlap. OD path observed proportions are obtained by aggregating OD paths showing a path size less a given threshold that has been set initially to 0.2.

A general form of the deterministic part of the utility function has been initially stated. It contains as explanatory path variables mean observed travel time, travel time standard deviation, mean trip length, trip length standard deviation, the proportion of main streets and a logarithmic transformation of mean path size. Estimation of route choice parameters is addressed using mlogit package in R.

A General Framework for route choice analysis based on GPS traces can be summarized in the following steps:

- Filtering internal trips to the study area, non-fleet provider type and reasonable trip length.
- Filtering outliers on trip travel times and trip mean speed.
- Filtering trips with latency over 3 min.
- Map matching of waypoints to INRIX XD-Segments.
- Classification of trips according to day-type and hour based on speed profiles leading to the identification of clusters of trips according to similar congestion.
- Aggregation of OD TAZ trips to OD municipality-district trips within the same congestion cluster.
- Selecting OD pairs with the largest number of trips (10 OD pairs) within the same congestion cluster.
- Calculation of path size to identify OD trips with common paths. Define OD path choice set.
- Calculation of OD path choice set variables and observed OD path proportions.
- Estimation of random utility model for OD path choice.

4. Preliminary results and discussion

Datasets are processed with RStudio 1.1.463 (2018) and R version 3.5.2 (2018) (R Core Team 2019) on a Windows 10 x86_64-w64-mingw32/x64 (64-bit) platform with the 30 GB RAM required for data exploitation. The data was prepared and the resulting working samples revealed no significant number of trips when considering spatial aggregation in TAZ for OD trips. No OD pairs satisfy a reasonable number of trips, although congestion level was not taken into account. Once spatial aggregation to the district in the municipality was considered and congestion level was controlled for in order to guarantee the stability of traffic conditions, few OD pair candidates remain for the path choice modeling exercise.

5. Conclusions

Map-matching algorithms between waypoints and XD Segments proved to be exceptionally difficult. There are many ways to design a map-matching algorithm in which each algorithm has advantages and disadvantages. The most complex version of the map matching problem (Greenfeld 2002) applies to the current situation in which the only information available is the network structure in the form of XD Segments and the positions of observations. Map-matching using PostGIS does provide acceptable, but not optimal results and map-matching algorithms are currently being implemented based on proposals by (Jagadeesh and Srikanthan 2016; Jagadeesh, Srikanthan, and Zhang 2004).

Although, GPS traces for one month contain a large amount of trip and waypoint registers, the proposed methodology drastically reduces trip samples and OD choice set of paths are usually very small, and hence a subset of the most observed OD pairs has to be selected for path size and choice set identification. In all selected OD pairs, less than 5-10 *different paths* (according to path size threshold parameter) are calculated. The researchers have found to be extremely difficult to fulfill the ambitious initial objectives of this paper, due to data preparation required time and limited sample size. GPS tracking data for a longer period of time is required. GPS tracking data is an innovative source that must be explored further, but traditional survey sources prove to deliver remarkable benefits in terms of direct use, although there are disadvantages regarding cost and preparing the data collection.

Acknowledgements

This research was funded by TRA2016-76914-C3-1-P Spanish R+D Programs and by Secretaria d'Universitats-i-Recerca-Generalitat de Catalunya- 2017-SGR- 1749 and Industrial PhD Program 2017 DI 041.

References

Anon. 2018. "PostGIS — Spatial and Geographic Objects for PostgreSQL."

Antoniou, C., L. Dimitriou, and F. Pereira. 2019. *Mobility Patterns, Big Data and Transport Analytics : Tools and Applications for Modeling*.

Ben-Akiva, M. and M. Bierlaire. 1999. "Discrete Choice Methods and Their Applications to Short Term Travel Decisions." Pp. 5–33 in. Springer, Boston, MA.

Ben-Akiva, M. and S. Lerman. 1985. *Discrete Choice Analysis : Theory and Application to Travel Demand*. MIT Press.

Cascetta, E., F. Russo, and A. Vitetta. 1997. "Stochastic User Equilibrium Assignment with Explicit Path Enumeration: Comparison of Models and Algorithms." *IFAC Proceedings Volumes* 30(8):1031–37.

Cerqueira, Vitor, Luis Moreira-Matias, Jihed Khiari, and Hans van Lint. 2018. "On Evaluating Floating Car Data Quality for Knowledge Discovery." *IEEE Transactions on Intelligent Transportation Systems* 19(11):3749–60.

Chen, C., J. Ma, Y. Susilo, Y. Liu, and M. Wang. 2016. "The Promises of Big Data and Small Data for Travel Behavior (Aka Human Mobility) Analysis." *Transportation Research Part C: Emerging Technologies* 68:285–99.

Chodrow, Philip S., Zeyad al-Awwad, Shan Jiang, and Marta C. González. 2016. "Demand and Congestion in Multiplex Transportation Networks" edited by Y. Moreno. *PLOS ONE* 11(9):e0161738.

Frejinger, E., M. Bierlaire, and M. Ben-Akiva. 2009. "Sampling of Alternatives for Route Choice Modeling." *Transportation Research Part B: Methodological* 43(10):984–94.

Gibert, K., M. Sànchez–Marrè, and J. Izquierdo. 2016. "A Survey on Pre-Processing Techniques: Relevant Issues in the Context of Environmental Data Mining" edited by Karina Gibert. *AI Communications* 29(6):627–63.

Greenfeld, J. 2002. "Matching GPS Observations to Locations on a Digital Map." in *81th Annual Meeting of the Transportation Research Board*.

Hensher, D. A. and W. H. Greene. 2003. "The Mixed Logit Model: The State of Practice." *Transportation* 30(2):133–76.

Jagadeesh, G. R. and T. Srikanthan. 2016. "Heuristic Optimization for High-Speed Low-Latency Online Map Matching with Probabilistic Sequence Models." Pp. 2565–70 in *19th International Conference on Intelligent Transportation Systems (ITSC)*.

Jagadeesh, G. R., T. Srikanthan, and X. D. Zhang. 2004. "A Map Matching Method for GPS Based Real-Time Vehicle Location." *Journal of Navigation* 57(3):429–40.

Montero, L., X. Ros-Roca, R. Herranz, and J. Barceló. 2019. "Fusing Mobile Phone Data with Other Data Sources to Generate Input OD Matrices for Transport Models." *Transportation Research Procedia* 37:417–24.

Montini, L., C. Antoniou, and K. W. Axhausen. 2017. "Route and Mode Choice Models Using GPS Data." *TRB Annual Meeting Online* 1204:17–03082.

de Moraes-Ramos, G., W. Daamen, and S. Hoogendoorn. 2013. "Modelling Travellers' Heterogeneous Route Choice Behaviour as Prospect Maximizers." *Journal of Choice Modelling* 6:17–33.

R Core Team. 2019. "R: A Language and Environment for Statistical Computing."

Shen, L. and P. Stopher. 2014. "Review of GPS Travel Survey and GPS Data-Processing Methods." *Transport Reviews* 34(3):316–34.

Toole, J., S. Colak, B. Sturt, L. Alexander, A. Evsukoff, and M. González. 2015. "The Path Most Traveled: Travel Demand Estimation Using Big Data Resources." *Transportation Research Part C: Emerging Technologies* 58:162–77.