Journal of Choice Modelling 42 (2022) 100337

Contents lists available at ScienceDirect

Journal of Choice Modelling

journal homepage: www.elsevier.com/locate/jocm

Long-distance mode choice model estimation using mobile phone network data

Angelica Andersson[a,b,*], Leonid Engelson[b], Maria Börjesson[a,b], Andrew Daly[c], Ida Kristoffersson[a]

[a] *VTI Swedish National Road and Transport Research Institute, Sweden*
[b] *Linköping University, Sweden*
[c] *ITS, University of Leeds, United Kingdom*

ARTICLE INFO

Keywords:
Demand model
Mode choice
Mobile phone network data
Travel behaviour
Long-distance travel

ABSTRACT

In this paper we develop two methods for the use of mobile phone data to support the estimation of long-distance mode choice models. Both methods are based on logit formulations in which we define likelihood functions and use maximum likelihood estimation. Mobile phone data consists of information about a sequence of antennae that have detected each phone, so the mode choice is not actually observed. In the first trip-based method, the mode of each trip is inferred by a separate procedure, and the estimation process is then straightforward. However, since it is not always possible to determine the mode choice with certainty (although it is possible in the majority of cases), this method might give biased results. In our second antenna-based method we therefore base the likelihood function on the sequences of antennae that have detected the phones. The estimation aims at finding a parameter vector in the mode choice model that would explain the observed sequences best. The main challenge with the antenna-based method is the need for detailed resolution of the available data. In this paper we show the derivation of the two methods, that they coincide in case of certainty about the chosen mode and discuss the validity of assumptions and their advantages and disadvantages. Furthermore, we apply the first trip-based method to empirical data and compare the results of two different ways of implementing it.

1. Introduction

Transport policy development requires demand forecasts, so that different scenarios can be appraised. Travel demand models have traditionally been estimated based on national travel surveys (NTS). Low response rates have however become a major issue of NTS data in recent decades (Prelipcean et al., 2018) and this may inflict sample bias. There is for example a risk that travellers with high value of time are less likely to respond to the survey (Stopher and Greaves, 2007). Data collection based on GPS tracking has also become popular, because it reduces response errors and can include route choice information which is usually absent in NTS. However, GPS tracking surveys suffer from even lower response rates than NTS (Indebetou and Alexander, 2018). Due to these concerns with low response rates, and the high costs of data collection, more attention has been directed towards mobile phone data in travel demand modelling. Such data is purely passive and thus does not depend on active response from travellers. The number of collected observations can also be very large at a low cost for the operator. However, the characteristics, error structure and format of mobile phone

* Corresponding author. VTI Swedish National Road and Transport Research Institute, Sweden.
E-mail address: angelica.andersson@vti.se (A. Andersson).

https://doi.org/10.1016/j.jocm.2021.100337
Received 30 September 2020; Received in revised form 22 September 2021; Accepted 25 November 2021
Available online 1 December 2021

data is fundamentally different from NTS or GPS tracking data. The purpose of this paper is therefore the development of theoretical model formulations for a passive mobile phone dataset, and the validation of key aspects of the developed theory using real-world data.

A major advantage of the mobile network data source is not only that it is passive, but also that it covers almost the entire population in developed countries. For example, 99% of Swedes over the age of 10 have their own cell phone (Internetstiftelsen, 2019). Hence, as long as the chosen mobile phone operator whose data is used does not focus on a specific customer segment, the data has a high chance of being representative of the population.

So far research has mainly used mobile phone data for development of origin-destination-matrices (Alexander et al., 2015; Bekhor et al., 2013; Calabrese et al., 2011; Gundlegård et al., 2016; Tolouei et al., 2017; Toole et al., 2015) and also for cross-validation of transport models (Wu et al., 2018). Dypvik Landmark et al. (2021) assess the quality and robustness of mobile network data by comparing it with several other data sources. The authors conclude that mobile phone data is a worthwhile resource for transport-related applications, in particular for the case of long-distance trips, as the quality of the data decreases with a finer zonal resolution. To our knowledge, no prior studies have estimated mode choice models based solely on mobile phone network data. Bwambale et al. (2019) used mobile phone network data to model long-distance route choice and Bwambale et al. (2020) demonstrated the feasibility of a joint modelling framework for mobile phone and survey data in the case of trip generation, also commenting on the potential to apply similar models in the context of mode choice. Bwambale et al. (2019b) also used a combination of GPS and mobile phone network data to model departure time. To describe how activity patterns change over time and space, Diao et al. (2016) estimated a model for participation in different activities based on a travel survey and then applied it on mobile phone data so that the extent and variation of participation in activities at different times and places could be illustrated for an entire city. Two authors have used mobile phone data fused with survey data to define and estimate travel demand models. Janzen (2019) defined and estimated an activity-based demand model on a synthetic population. The synthetic population was based on survey data that was scaled to match the number of trips between different origin-destination pairs (OD pairs) in mobile phone data using histogram matching. Brederode et al. (2019) used a multi-proportional gravity model to fuse mobile phone and survey data into a “common operational picture” of the total travel demand before parameter estimation. Neither Janzen (2019) nor Brederode et al. (2019), considered uncertainty in chosen mode.

A key issue in the modelling of mode choice based on mobile phone data is that the mode choice is not directly observed, which it is in data traditionally used for mode choice models. Instead, only a sequence of signals from the different antennae are observed. This means that mode (and destination) choice model formulations must be adapted to take these features of the data into account. In this paper we develop two theoretical methods for the use of mobile phone data to support the estimation of long-distance mode choice models, both based on the logit model (McFadden, 1974) and maximum likelihood estimation (MLE) framework (Edwards, 1992): a “trip-based” method and an “antenna-based” method. We also estimate a mode choice model empirically by applying and comparing two approaches of the trip-based method.

The first, trip-based, method assumes that the trips and their characteristics (origin, destination, and chosen main mode, etc) are already identified from the raw data consisting of antenna connections. Breyer et al. (2021) develop methods for making such trip identification. They apply their methods to Swedish mobile phone network data. In this paper we use a data set generated by one of their methods, where a set of long-distance trips by air, road and rail are identified, to define and estimate a mode choice model applying our trip-based method. A complication when defining and estimating a mode choice model on this type of data is that the chosen mode cannot always be identified with certainty. Instead, the methods identify the chosen mode up to a probability, i.e., they report the probability that a mode, say rail, was chosen as main mode. This probability will be referred to as the identification probability. Hence, the choice outcome is not known with certainty. Even if the techniques for logit estimation have been developed to deal with measurement errors in the explanatory variables (Guerrero et al., 2020; Varela et al., 2018; Walker et al., 2010), it has rarely been developed for dealing with uncertainty about the choice outcome. Note, however, that even if the issue of uncertain choice outcomes has rarely been investigated for survey data, this is not to say that there are no reporting errors in the chosen mode stated in the surveys. Modellers have however assumed that the observed choice does not have a systematic bias. We emphasise that Breyer et al. do not define or estimate any mode or destination choice model that could be used in forecasting, which we do in the present paper, but aim only at identifying trips and their characteristics from mobile phone data, with a special focus on the chosen main mode.

In our trip-based method, we deal with the uncertainty in the observation of the chosen mode in two different ways. One is by assuming that the chosen mode is the mode for which the identification probability is the largest. The other way is to treat each observed trip as representative of several trips with identical characteristics except that the chosen mode is different. We then assume that the identification probabilities represent modal shares and estimate the model as if we had data where each observation was an aggregate over several observations. In the random utility theory literature, such aggregated data is typically applied in combination with a weighted likelihood function (Manski and Lerman, 1977), where the weights correspond to the number of observations that each aggregated observation represents. In our case, however, the weights are all set to unity since all observations still only represent one trip.

The second, antenna-based method, is instead formulated to use the sequences of antennae that have detected the phones. This model incorporates the chosen mode identification in the model definition in a consistent way. Our model definition is inspired by Bierlaire and Frejinger (2008) who estimate a route choice model based on GPS data. We are not able to estimate any mode choice model empirically applying the antenna-based method, since we do not have access to the required raw data. However, we still think it is useful as a theoretical benchmark for the sake of discussion. Also, it could be that other mobile phone network operators are more able to share data of this type than the one we have been in contact with. In that case our theoretical derivation can be used as a blueprint for the use of mobile phone network data to estimate mode choice demand models.

A problem with mobile phone data that we do not address in the present paper is that information such as socio-economic characteristics, party size and the purpose of the trip is not collected. It can therefore be useful to combine this data with NTS data in the estimation of demand models (Chen et al., 2016). The following sections will focus on the formulation of likelihood for the passive dataset (i.e., mobile phone data) into a format that later can be used in combined estimation with active data such as NTS. Section 2 describes the structure of mobile phone data and Section 3 contains the derivations of the two proposed methods. In Section 4 the "trip-based" model is empirically estimated on data from Breyer et al. (2021), and in Section 5 concluding remarks are made.

2. Structure of mobile phone data

2.1. *Features of mobile phone data*

The characteristics of mobile phone data are fundamentally different from those of travel survey data. The data consists of the sequences of antennae that have detected the mobile phones, so the mode is not directly observed. A smartphone sends and receives signals to an antenna every few minutes. These signals can be a phone call, sending or receiving a text message, actively using the internet, apps using internet in the background, or the phone operating system working in the background (Gundlegård, 2018). For each signal sent or received, the operator saves subscriber id, cell id and timestamps. The saved information about each such signal is referred to as a "data point" in this paper, and the information is generally referred to as xDR data.[1] Even though mobile phones have been used for decades, the emergence and growth in the number of smartphones has increased the number of signals sent and received per time unit by mobile phones and antennae. This means that mobile phone data now have an even greater opportunity to provide insights into human mobility patterns.

The probability of connecting to a specific antenna mainly depends on the proximity between phone and antenna. There are many factors and complications to consider when using mobile phone data to detect and model travel behaviour. First, each mobile network antenna has a certain coverage area. Within urban areas antennae can be located just a couple of hundred meters apart, while in rural areas antennae are more sparsely located (in the size order of 10 km apart) and there can be areas without coverage. Antennae in urban areas typically have a mixture of different coverage areas, such that some antennae can be found within the coverage area of another antenna (Gundlegård, 2018). This means that the position of the phone when data is sent, the position of the surrounding antennae, and the local topography, may affect which antennae a phone connects to during a trip. Other factors affecting the probability of connecting to an antenna includes the height of nearby buildings (affecting the reception quality from different directions), the current network load of different antennae, and the type of antenna the phone is configured to prefer (GSM, WCDMA, LTE or 5G) (Gundlegård, 2018). In essence these factors imply that two trips using the same mode, on the same route, between the same OD pair may connect to different sequences of antennae.

When using mobile phone network data for travel demand estimation purposes it is crucial to fulfil legal and ethical requirements for privacy protection. No information that can connect a mobile phone number or any other personal identifier (an explicit identifier) with a series of connections to mobile phone antennae can leave the mobile phone operator servers, if the mobile phone operator is to comply with EU privacy laws (GDPR). It is also necessary to make sure that no individual can be identified if the mobile data sent out from the mobile phone operator servers is used in combination with another data source (also called a quasi-identifier). Badu-Marfo et al. (2019) review anonymisation operations and techniques within the context of big transportation data. They conclude that the three most common anonymisation operations are generalisation, i.e. replacing precise information with a taxonomy of its parent value, suppression, i.e. simply removing some attribute, and perturbation, i.e. replacing some of the values of the individual attributes while maintaining the aggregate characteristics of the dataset. For the sake of this project, generalisation (of time of day, position and length of stay) and suppression (of any explicit identifiers) has been used before the data leaves the mobile phone operator servers.

Because of the above concerns, some aggregations of mobile phone data are easier to export from a mobile phone operator than others while still ensuring privacy law compliance. For instance, a table of trip observations where individuals have been hidden by the methods mentioned above is possible to export, in contrast to a full sequence of antenna connections, which would probably have to be processed on the servers of the mobile phone operator. The problem with trip-based methods is that there might be aggregation errors when going from antenna level to trip level.

To summarise, mobile phone data would ideally (for a transport modeller) come in the format of sequences of data points. A datapoint includes a timestamp and antenna position for a signal sent or received from a mobile phone. However, in some circumstances, mobile data might only be accessible to the modeller in some aggregated format such as a table of trip observations. The trip-based format and the antenna-based format are both different from the format of travel survey data. In section 3 we therefore develop methods to estimate demand models based on the two data formats: a trip-based method and an antenna-based method. However, before we are ready to define the methods, we give a short example of how the table of trip observations can be derived from the original data, i.e., the sequences of antennae that detected the phone.

2.2. *Features specific to the trip-based method*

The raw data consists of data points. To be able to estimate a mode choice model, origin, destination, start time and chosen mode must first be identified for each trip. Breyer et al. (2021) develop several methods for such trip and mode identification from the

[1] xDR stands for Detail Record of call, text or internet data, as opposed to the traditional CDR or Call Detail Record data, which only contain a Detail Record of calls and text messages.

antennae observations, and then apply the methods to Swedish data. The mode choice model formulation defined in Section 3.2, assumes that the data has the format produced by one of the methods developed by Breyer et al., called the route/antenna method, described in this section. The format of the resulting data set is demonstrated using example data in Table 1. Data of this format will also be used in the empirical estimation of the mode choice model in Section 4.

The mode identification method in Breyer et al. (2021) classifies trips as either rail or road. The method uses the distance between each antenna detecting the phone and the routes by different modes to find measures of mode identification probabilities of each mode. First, the sequence of antennae S_t corresponding to a trip, the trip origin and the trip destination are identified. Then, Open Trip Planner ("OpenTripPlanner," 2020) is used to generate routes between trip origin and trip destination by mode $m \in M = \{\text{rail, road}\}$. For road, only one route $R_{t,road}$, the route with the minimum generalised cost (measured as shortest travel time), is generated. For rail, a set of up to three possible routes with departure times close to the start time of the trip, are generated, using the train timetable. The route $R_{t,rail}$ is then the route that minimises the average line distance between the route and the sequence of antennae. The identification probability $q_{tm} = P(m|S_t)$ of mode m for trip t, is calculated in Breyer et al. (2021) as inversely proportional to the average distance between antennae in the registered sequence S_t and the route:

$$q_{tm} = \frac{\left(\frac{1}{N_t}\sum_{a\in S_t} dist(a, R_{tm})\right)^{-1}}{\sum_{k\in \mathrm{M}}\left(\frac{1}{N_t}\sum_{a\in S_t} dist(a, R_{tk})\right)^{-1}} \tag{1}$$

where $N_t = |S_t|$ is the total number of antennae which the phone has connected to during trip t, and $dist(a, R_{tm})$ is the shortest line distance between antenna a and any point on the route R_{tm}. Note that Equation (1) is not an approximation of a well-defined conditional probability; it is rather a possible way to quantify a likelihood of the used mode based on comparison of distances from the antennae that detected the phone to plausible routes for different modes.

Section 4 exemplifies how the identification probabilities can either be used directly as the choice outcome of the main mode in the model estimation, or by using the mode with the largest identification probability as the chosen mode.

We aim at estimating a mode choice model for long-distance travel that include the main modes rail, air, car, and bus. The methods developed by Breyer et al. (2021) identify only the main modes rail or road. After publishing their paper, Breyer et al. have also developed a method to identify air trips.[2] In the data that they have provided for this paper, identification of air trips is done in a separate step before the route/antenna method is applied. That is, only trips that were identified as non-air trips in the first step are later subject to the route/antenna method. The identification of car and bus, respectively, for trips identified as road trips by Breyer et al. can be handled in different ways within the mode choice model estimation, which are described in Section 3.6.

The antenna coordinates for origin and destination of the identified trip are then assigned to zones, taken to be the zones of the model for national long-distance trips used by the Swedish Transport Administration. These zones cover the whole country, in total there are 682 zones, which means that each zone on average covers about 770 km^2 and 15 000 inhabitants, although the sizes and populations of the zones vary depending on land use density. The large geographic size of the zones is an advantage in the sense that a zone usually covers a greater area than that of the coverage area of an antenna, but a disadvantage in the sense that it will not be possible to accurately derive any socio-economic variables based on origin or destination zone. It is assumed that the origin zone and destination zone is known with certainty. This assumption may be violated if a trip end is near the border between two zones. However, it is not a problem in most cases since the long-distance zones of the Swedish Traffic Administration are larger than the reception area of a mobile network antenna.

Table 1 illustrates the format of the data used in the model estimation of Section 4 along with example data containing the timing of the trip (weekday or weekend, peak or off-peak), origin zone, destination zone, identification probability for road, rail and air, and most probable mode.

3. Analysis methods for mobile phone data

We proceed by formulating two different methods for estimating a long-distance mode choice model on mobile phone data. The first estimation method is based on data derived by a procedure such as the one described in Section 2.2. The second method, developed for comparison, would use the more fundamental data, i.e., the sequence of antennae that has detected each phone, so the mode choice is not actually observed.

3.1. Formulation of likelihood maximisation problem

Maximum likelihood estimation (MLE) is used in the two model formulations, i.e., we maximise the likelihood function so that the observed data is most probable under the assumed statistical model. We want to find a parameter vector β that would best explain the observations g_t (in our trip-based method g_t is the identified mode of trip t and in our antenna-based method g_t is the vector S_t of

[2] Identification of air trips is performed in two steps for a given trip: 1. Pick out any airports closer than 10 km to any connected antenna; 2. If the speed between any pair of airports identified in step 1 is over 200 km/h and the distance between the two airports is larger than 200 km, the trip is classified as an air trip. Thus, the identification probability of air is always 0 or 1.

Table 1
Illustrative input data format for the trip-based method containing example data.

Timing of trip	Origin	Destination	Identification probability road	Identification probability rail	Identification probability air	Largest identification probability mode
Weekday, peak	1	3	1	0	0	Road
Weekday, off-peak	1	2	0.89	0.11	0	Road
Weekday, peak	1	2	0.05	0.95	0	Rail
Weekend, peak	1	2	0	0	1	Air
Weekday, peak	5	3	0.03	0.97	0	Rail
Weekend, off-peak	5	3	0.49	0.51	0	Rail

antenna positions within the trip t). The likelihood is the probability of observing g_t given β, i.e., $L(g_t,\beta) = P(g_t|\beta)$. When the observation includes exogenous variables x, the likelihood becomes $L(g_t,\beta) = P(g_t|\beta,x)$. Finally, if T is the set of all identified trips t, each with origin o_t, destination d_t, exogenous variables x_t and choice g_t, the overall likelihood becomes

$$L(\beta) = \prod_{t\in T} P(g_t|\beta, o_t, d_t, x_t). \tag{2}$$

Maximisation of this expression is equivalent to

$$\max_{\beta} \ln L(\beta) = \max_{\beta} \sum_{t\in T} \ln P(g_t|\beta, o_t, d_t, x_t). \tag{3}$$

A general problem when applying mobile phone network data to estimate behavioural models is that the data include few explanatory variables. For this reason, the model could be subject to endogeneity problems caused by omitted variables. In the case of transport demand models, the data may also be subject to measurement errors regarding the traffic supply variables, which is an additional source of endogeneity. The latter is not unique for transport demand models estimated on mobile phone data. In fact, both Guerrero et al. (2020) and Varela et al. (2018) show that more conventional mode choice models are subject to endogeneity bias, affecting costs more than times, and therefore implying that the value of time is overestimated.

However, endogeneity problems in choice models can be addressed by several approaches. Latent variables can be used to correct for measurement errors (Varela et al., 2018; Walker et al., 2010) and recently Guerrero et al. (2020) further developed Heckman's (1978) control function, or instrumental variable, approach for mode choice models to account for endogeneity. Guerrero et al. (2020) developed a method to make forecasts by applying these models. Hence, even if endogeneity problems could arise, they could in many cases be handled, as long as a valid and relevant instrument can be identified.

3.2. Trip-based observation

In the data used for model estimation in the present paper, generated by the process described in Section 2.2, the chosen main mode is not always known with certainty. As the standard logit formulation assumes that the chosen alternative is known with certainty, it is necessary to make some assumption regarding the chosen mode in the estimation.

One approach would be to assume that each observed trip is representative for the mode choice of a number of trips, distributed over the alternative modes according to the chosen mode identification probabilities as given by Equation (1). According to Equation (3), maximisation of the log likelihood is then formulated as

$$\max_{\beta} \ln L(\beta) = \max_{\beta} \sum_{t\in T} \sum_{m\in \mathrm{M}} q_{tm} \ \ln P(m|\beta, o_t, d_t, x_t) \tag{4}$$

where q_{tm} is the identification probability of mode m for trip t as given in Equation (1), x_t is the exogenous variables for trip t, m runs over the set of modes, and β contains the parameters to be estimated. The mode choice probabilities are modelled assuming the logit model

$$P(m|\beta, o_t, d_t, x_t) = \frac{\exp V_m(\beta, o_t, d_t, x_t)}{\sum_k \exp V_k(\beta, o_t, d_t, x_t)} \tag{5}$$

where V_m is the observed part of the utility function for mode m.

A second approach would be to take the mode with the largest identification probability, defined in Equation (1), as the chosen mode. That is, in equation (4) $q_{tm} = 1$ for the largest identification probability mode and $q_{tm} = 0$ for the other modes. We will apply both of these approaches in the model estimation of Section 4.

3.3. Antenna-based observation

As mentioned, the challenge with estimation of a mode choice model based on mobile phone data is that the mode of each trip is not always known with certainty, i.e., what is observed is not directly the chosen mode but the sequence of antennae that detected the phone. Therefore, in the theoretically consistent model formulation developed in this section, the observation vector g is set to be the sequence of antennae.

Let A be the universal set of all antennae used by a mobile phone operator in the study area. The main unit of analysis is a trip $t \in T$. This method is based on the observation defined as a combination (x_t, o_t, d_t, S_t) of a vector x_t of exogenous variables, trip origin o_t, trip destination d_t, and a set $S_t \subset A$ of antennae that detected the phone during the trip. The exogenous variables x_t include trip attributes apart from origin and destination. It is assumed that the origin and the destination of each trip are known with certainty. On the other hand, the trip mode cannot always be determined with certainty, i.e., the same set of antennae may detect trips performed by different modes. Therefore, in order to calculate the probability that the phone has connected to a certain set of antennae during the trip, we need to use conditional probabilities and to sum over different modes for each observed set of antennae. The likelihood of observation t then becomes

$$P\left(S_t|\beta, o_t, d_t, x_t) = \sum_{m\in M} p_{tm} P(m|\beta, o_t, d_t, x_t \right), \tag{6}$$

where M is the set of all considered modes, while $p_{tm} = P(S_t|o_t, d_t, m)$ is the probability that the set of antennae detecting the phone during the trip is S_t, given that the trip is performed from o_t to d_t by mode m; we assume that this probability is independent of x_t. Finally, $P(m|\beta, o_t, d_t, x_t)$ is the probability that a traveller characterised by x_t and travelling from o_t to d_t chooses mode m.

Note that definition of p_{tm} is "reverse" compared to the identification probability q_{tm} of mode given the set of antennae that is used in the trip-based method in section 3.2. These conditional probabilities are not proportional but can be expressed via each other using the Bayes theorem.

Which antennae detect the phone during the trip depends not only on the mode but also on the route taken by the traveller. Therefore, in order to calculate the probability p_{tm} we have to assume a route choice model. Let $R_{o,d,m}$ be the set of routes connecting o and d by mode $m \in M$.

Assumption 1. Assume there is a route choice model available that for every o, d, m and $r \in R_{o,d,m}$ determines the probability $P(r|o, d, m)$.

Now the probability that the set of antennae that detected the phone during the trip is S_t given that the trip went from o_t to d_t by mode m can be expressed as

$$p_{tm} = \sum_{r\in R_{o_t, d_t, m}} P(S_t|r) P(r|o_t, d_t, m) \tag{7}$$

where $P(S_t|r)$ is the probability that the set of antennae detecting a trip along route r is exactly S_t. Substituting (7) into (6) we obtain the total probability of the set of antennae S_t given the exogenous variables x_t and parameters β as

$$P\left(S_t|\beta, x) = \sum_{m\in M} \sum_{r\in R_{o_t, d_t, m}} P(S_t|r \right) P(r|o_t, d_t, m) P(m|\beta, o_t, d_t, x_t). \tag{8}$$

It remains to determine the conditional probability of the set of antennae S_t given route r. As a first approximation, we can make the assumption:

Assumption 2. The probability that a trip is detected by an antenna is a decreasing function of the distance between the route and the antenna, for example:

$$P(a|r) = p_0 e^{\alpha \cdot dist(r,a)} \tag{9}$$

where p_0 is the probability that the trip will be detected by an antenna located directly on the route, $dist(r, a)$ is the minimal distance between the route r and the antenna a, and α is a negative deterrence parameter. It might be efficient to assume $P(a|r) = 0$ for longer distances.

In order to calculate the probability of a set of antennae we need to make an additional assumption:

Assumption 3. The connections to different antennae are independent given a route.

Based on this assumption one can calculate the probability for a set of antennae as

$$P(S_t|r) = \prod_{a\in S_t} P(a|r) \prod_{a\in A\setminus S_t} [1 - P(a|r)], \tag{10}$$

which is the probability that every antenna in S_t and no other antennae detected the trip.

The final likelihood maximisation problem is thus

$$\max_{\beta}\ln L(\beta) = \max_{\beta}\sum_{t\in T}\ln\sum_{m\in M}P(m|\beta, o_t, d_t, x_t)\left[\sum_{r\in R_{o_t,d_t,m}} P(S_t|r)P(r|o_t, d_t, m)\right] \tag{11}$$

where $P(S_t|r)$ is calculated by (10) using (9).

For the estimation of the parameters β one needs the origin, the destination, the exogenous variables, and the value of the expression in the square parenthesis, i.e., p_{tm} for each trip t and each mode m. In the expression for p_{tm}, the first term, $P(S_t|r)$, needs to be provided by a mobile phone network operator calculated using Equation (10) and their knowledge of antenna coverage areas, and possibly also other factors such as the current network load at the time of connection. The second term $P(r|o_t, d_t, m)$, can be approximated using a route choice model. Calculation of p_{tm} using the set of antennae detecting the trip can be done at the premises of the mobile network operator as a data preparation step, preserving the privacy of the travellers.

3.4. The case of certain identification of the chosen mode

The trip-based and antenna-based methods are equivalent if the chosen mode is identified correctly and with certainty: the two methods result in the same log likelihood function. To show this, consider first the antenna-based method. The mode of the trip is uniquely determined by the set of antennae detecting the trip. For any trip t, the probability that the set of antennae that detected the phone during the trip is S_t given that the trip went from o_t to d_t by mode m is $p_{tm} > 0$ for a unique mode which we denote m_t, i.e. $p_{tm} = P(S_t|o_t, d_t, m) = 0$ for any $m \neq m_t$. The likelihood of Equation (6) then becomes:

$$P(S_t|\beta, o_t, d_t, x_t) = \sum_{m\in M} p_{tm}P(m|\beta, o_t, d_t, x_t) = p_{tm_t}P(m_t|\beta, o_t, d_t, x_t), \tag{12}$$

and the maximisation in Equation (11) reduces to $\max_{\beta}\ln L(\beta)$ where:

$$\ln L(\beta) = \sum_{t\in T}\ln\sum_{m\in M}P(m|\beta, o_t, d_t, x_t)p_{tm} = \sum_{t\in T}\ln\left[P(m_t|\beta, o_t, d_t, x_t)p_{tm_t}\right] = \sum_{t\in T}\ln P(m_t|\beta, o_t, d_t, x_t) + \sum_{t\in T}\ln p_{tm_t} \tag{13}$$

which means that the maximisation is equivalent to

$$\max_{\beta}\sum_{t\in T}\ln P(m_t|\beta, o_t, d_t, x_t). \tag{14}$$

Turning now to the trip-based method by applying the Bayes theorem for any $m \neq m_t$, we have $q_{tm} = P(m|S_t, o_t, d_t) = \frac{P(S_t|o_t,d_t,m)P(m|o_t,d_t)}{P(S_t|o_t,d_t)} = 0$, which implies $q_{tm_t} = 1$ (this only holds in the extreme case of correct and certain mode identification; see however section 3.5 below considering the case of incorrect identification). Then, using $q_{tm_t} = 1$ and $q_{tm} = 0$ for any $m \neq m_t$ in Equation (4) we obtain $\max_{\beta}\ln L(\beta) = \max_{\beta}\sum_{t\in T}\ln P(m_t|\beta, o_t, d_t, x_t)$, which is the same as in the antenna-based method under the assumption of certainty in the mode choice classification (Equation (14)).

3.5. The case of erroneous identification of the chosen mode

A problem occurs when a significant fraction of the trips uses a suboptimal road route, then the mode identification probabilities can be subject to errors. Fig. 1 describes a network when such errors are present. There is one railway and two roads, one with a low generalised cost (optimal road route) and one with a higher generalised cost (suboptimal road route). The suboptimal road route runs along the railway while the optimal road route runs far away. Remember that in the mode identification method described in Section 2.2 the mode probabilities are calculated based on the distance between the antennae detecting the phone and the optimal road route and the closest rail route, ignoring suboptimal road routes. Assume that there are just two antennae: one close to each of the two roads, and their coverage areas do not overlap so that the calculated $q_{tm}{}'s$ are close to 1 or 0. If a phone is carried by a traveller along the suboptimal road route, it will incorrectly be identified as a rail trip with nearly full certainty in the mode identification procedure. The described errors will only be present in the trip-based method and the magnitude of the error thus caused will depend on the road and rail networks and the flows on these.

While we do not have a measure of accuracy for the full national dataset used as input in this paper, Breyer et al. (2021) tested their method on manually annotated trips between the cities Norrköping and Linköping, and found that the route antenna method used to generate the input data for this paper had an accuracy of 95.5%. That is, for 95.5% of the observations the traveller in the validation dataset of 510 trips for which mode choice was known with certainty, had actually chosen the most probable mode. Furthermore, also travel surveys are likely to be subject to errors (Houston et al., 2014; Stopher et al., 2007), so that the dependent variable might also be incorrect in that case. Even though it is not necessarily the case that the accuracy of the validation dataset is exactly the same as for the full dataset, the figure still gives an indication of the expected quality of the input data.

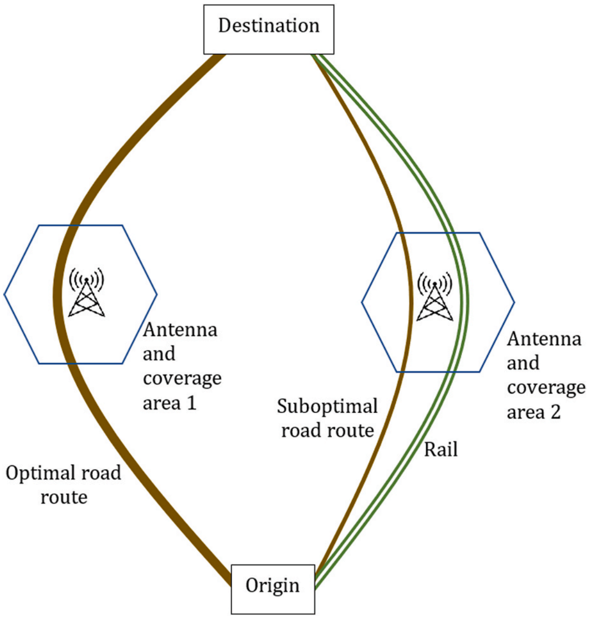

Fig. 1. Assumed geography.

3.6. *Handling the road mode*

Three options may be considered to handle the fact that car trips and bus trips are both represented as “road” in the passive dataset.[3] The first option leaves the choice to be assigned to “road” using an ‘average’ road utility function, for example:

$$V_{road} = ASC_{road} + \beta_{timeroad}(t_{car}(1-b) + t_{bus}b) + \beta_{cost}(c_{car}(1-b) + c_{bus}b) \tag{15}$$

Where β are parameters to be estimated, t denotes time, c denotes cost and b is the fraction of bus trips observed from a previous travel survey.

The second option is to randomly assign the mode to either the choice car or the choice bus, with probabilities of the two modes proportional to mode shares from a previous travel survey, i.e. using b as in equation (15).

The third option leaves the choice to be assigned only to road (similarly to option 1). Here, the joint probability of choosing the nest road and choosing mode i within the road nest (where $i \in \{\text{car, bus}\}$ and $k \in \{\text{road, train, air}\}$) is given by:

$$p_{road,i} = Pr(road)Pr(i|road) = \frac{\exp\ V^*_{road}}{\sum_k \exp V^*_k} \frac{\exp V_{road,i}}{\sum_j \exp V_{road,j}} \tag{16}$$

where $V^*_{road} = \theta \log \sum_j \exp\left(V_{road,j}/\theta\right)$, and $V_{road,j}$ is the utility for nest road and mode $j \in \{\text{car, bus}\}$. The parameter, $0 < \theta \leq 1$ is a structural coefficient determining the scale of choices at the car/bus-level compared to choices at the road nest level.

The selected option for handling the road mode will be the one that produces the best result, i.e. highest likelihood, on the dataset being used.

4. Application

In order to validate the trip-based method, two simple models are estimated. Work is ongoing to represent the road mode, trip purpose, access trips and additional explanatory variables in a suitable way in the model, but in order to validate the general method, a simple model is useful. In the first of the estimated models it is assumed that the mode identification is certain, while in the second model mode identification probabilities, as in Equation (1), are used as mode shares.

The specifications and assumptions used for the two estimations are otherwise the same. For these simple models it is assumed that all road trips are car trips; in fact, just 5–6% of long-distance road trips in Sweden are by bus (Berglund and Kristoffersson, 2020), so these models should give a good indication of the validity of the specifications. The utility functions are specified as:

[3] Breyer et al. (2021) initially considered the option of separating car and bus trips by speed at an earlier stage of the data processing, but it turned out that the time resolution was not high enough to accommodate this.

$$V_{car} = ASC_{car} + \beta_{cost} c_{car} + \beta_{time\ car} t_{car} \tag{17}$$

$$V_{rail} = ASC_{rail} + \beta_{cost} c_{rail} + \beta_{time\ rail} t_{rail} \tag{18}$$

$$V_{air} = ASC_{air} + \beta_{cost} c_{air} + \beta_{time\ air} t_{air} \tag{19}$$

where c denotes cost and t in-vehicle time for each respective mode; ASC denotes the alternative-specific constant for each mode. The estimations were performed on a dataset of roughly 92 000 trips sampled from one week in 2018. Supply data containing travel times and travel costs between different zones was provided by the Swedish Transport Administration from their current large-scale model of long-distance trips in Sweden. ASC_{car} is normalised to 0.

Table 2 shows the estimation results. The log likelihood values of the two models suggest that the specification using the largest identification probability as the chosen mode is the most successful in describing the data. However, as the discrete choice model specifications of the two approaches are identical (the difference lie in the format of the inputs of the models), the difference in likelihood values only implies that the mode with the highest identification probability gets predicted better than the other modes. All parameters except β_{cost} show either slightly better or much better accuracy in the largest identification probability as chosen mode model. β_{cost} show slightly better accuracy in the identification probabilities as mode shares case, although the t-ratios are of comparable size. Another reason to prefer the model that assumes that the mode with the highest identification probability is the chosen mode is that the calculation of the identification probabilities is somewhat arbitrary and could be changed by altering the exponent in Equation (1). Also, the identification probabilities are only applicable when distinguishing between rail and road modes, the air mode is identified using a separate method.

The values of time of all the modes are high, especially for air trips, even when accounting for the fact that the dataset contains a mix of private and business trips. This is to be expected as access trips are not accounted for in this simple model, and access trips typically comprise a significant share of total travel time for air trips. Time spent at the airport for security checks and mandatory check-in times are not included either. Also, the measurement of travel cost in the network data is subject to error, since travel costs from a large-scale model of long-distance trips in Sweden are used which has its limitations in representing travel costs and especially cost variation (Kristoffersson et al., 2020).

5. Conclusion

The trip-based method uses conventional multinomial logit MLE on a table of trips with mode identification probabilities as input. The estimation of the mode choice model is then straightforward. The key issue with this framework is that the dependent variable, the mode choice itself, is not known with certainty for all trips. There is not yet any developed method of consistently incorporating such uncertainties in a logit model. However, mode choice detection with full certainty for all observations is almost impossible in mobile phone network data. For this reason, all previous studies known to us on choice modelling based on such data use some form of our trip-based methods. The mode choice is not known with certainty in the antenna-based method either, but the antenna-based method is defined to handle the uncertainty in the mode choice in a consistent way.

However, applying the antenna-based method in practice would require an extremely intense collaboration between researchers and a mobile phone operator company. The reason for this lies in Equation (11). It states that the full optimisation of the parameters β needs to be performed with access to the probability of connecting to the specific antennae that the phone has connected to during the trip $P(S_t|r)$, and also the probability $P(r|o_t, d_t, m)$ (that can be approximated by a route choice model), along with the origin zone, destination zone and timing of trip. Hence, our antenna-based method should be viewed as the theoretically consistent way of estimating mode choice on mobile phone network data. It is therefore useful as a benchmark even if it might not be possible to apply it in practice. Note also that when the mode is identified with certainty the trip-based and the antenna-based methods are identical.

Table 2
Estimation assuming that the mode identification is certain (left) and estimation using the identification probabilities as mode shares (right), with t-values given within parentheses. Values of time converted from Swedish Krona to Euro using conversion rate of 10.

	Assuming that mode identification is certain	Mode identification probabilities as mode shares
Observations	92011	92011
Final log (L)	−63 130.104	−65 788.099
D.O.F.	6	6
ASC_{car}	0	0
ASC_{air}	−3.014 (−36.0)	−2.63 (−32.7)
ASC_{rail}	−0.7620 (−48.7)	−0.405 (−26.9)
β_{cost}	−0.00101 (−15.3)	−0.00112 (−17.1)
$\beta_{time\ air}$	−0.0164 (−16.3)	−0.0144 (−15.8)
$\beta_{time\ car}$	−0.00715 (−51.9)	−0.00571 (−42.2)
$\beta_{time\ rail}$	−0.00597 (−43.4)	−0.00537 (−40.0)
VoT_{air}	97.16 €/h	76.92 €/h
VoT_{car}	42.33 €/h	30.48 €/h
VoT_{rail}	35.32 €/h	28.67 €/h

In the derivation of the antenna-based method, some assumptions regarding the data generating process were made. It is relevant to discuss whether the necessary assumptions are valid. Assumption 1 holds if route shares for all routes in the network are known, which is unrealistic for real road networks. As an approximation one could use the results of a stochastic traffic assignment, although it is unclear how the limitations of the route generation algorithm and the route choice model would affect the estimation results. Assumptions 2 and 3 would hold if the signal strength depended only on the distance between the antenna and the phone. In practice, however, the signal strength at a given position is affected by buildings in the local surroundings and by the antenna's direction and range. Apart from the signal strength, the detection depends on current network load and the network type the phone is configured to use[4] (Gundlegård, 2018). Assumption 3 might be violated if the choice of antenna to connect to is affected by the previous connection (s). If the mobile phone is moving, the range of the currently connected antenna and the timing of when the mobile phone sends data will interact with the other factors and affect the choice of the next antenna (Gundlegård, 2018).

A key problem with our application of the trip-based method is that the mode choice probabilities might be systematically biased since the mode identification probabilities are calculated under the assumption that there is only one used road route per mode. However, the pre-processing method used to generate the identification probabilities indicated a high accuracy when tested on a smaller validation dataset: for 95.5% of the observations the traveller in the validation dataset (for which mode choice was known with certainty) had actually chosen the most probable mode. Furthermore, also travel surveys are likely to be subject to errors (Houston et al., 2014; Stopher et al., 2007), so that the dependent variable might be uncertain also in NTS data, although this problem has not been discussed by transport modellers in the literature.

The empirical estimation results of the trip-based method, indicates that it is feasible and provides intuitive results with high significance. Furthermore, any errors in the trip-based method are likely to be small, given the high accuracy of the identification probabilities when validated. Comparing the results of the versions of the two trip-based approaches, the approach that treats the mode with the highest identification probability as the chosen mode gives the best model fit.

The main contribution of this paper is the development of two theoretical model formulations and one empirical application for mode choice model estimation on a passive mobile phone dataset, which can later be used in combination with other data sources to make a joint estimation. In particular, long-distance mode choice models are important as decision support for developing high-speed railways and regulation of the flight communication markets. The two methods we present offer rational ways to work with mobile phone data which appear to be applicable in a wide range of circumstances.

Glossary

MLE	Maximum Likelihood estimation
NTS	National Travel Survey
OD matrix	A matrix of e.g. number of trips between origin zone (row) and destination zone (column)
xDR	Call and internet Detail Record

References

Alexander, L., Jiang, S., Murga, M., González, M.C., 2015. Origin–destination trips by purpose and time of day inferred from mobile phone data. Transport. Res. C Emerg. Technol. 58, 240–250.

Badu-Marfo, G., Farooq, B., Patterson, Z., 2019. A perspective on the challenges and opportunities for privacy-aware big transportation data. J. Big Data Anal. Transp. 1, 1–23.

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

Berglund, S., Kristoffersson, I., 2020. Anslutningsresor: en deskriptiv analys (Connection trips: a descriptive analysis) (No. 2020:3). Working Papers in TransportEconomics.

Bierlaire, M., Frejinger, E., 2008. Route choice modeling with network-free data. Transport. Res. C Emerg. Technol. 16, 187–198. https://doi.org/10.1016/j.trc.2007.07.007.

Brederode, L., Pots, M., Fransen, R., Brethouwer, J.-T., 2019. Big Data fusion and parametrization for strategic transport demand models. In: 2019 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS). Presented at the 2019 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), pp. 1–8. https://doi.org/10.1109/MTITS.2019.8883333.

Breyer, N., Gundlegård, D., Rydergren, C., 2021. Travel mode classification of intercity trips using cellular network data. Transp. Res. Procedia 52, 211–218.

Bwambale, A., Choudhury, C., Hess, S., 2019a. Modelling long-distance route choice using mobile phone call detail record data: a case study of Senegal. Transp. Transp. Sci. 15, 1543–1568. https://doi.org/10.1080/23249935.2019.1611970.

Bwambale, A., Choudhury, C.F., Hess, S., 2019b. Modelling departure time choice using mobile phone data. Transport. Res. Part Policy Pract. 130, 424–439. https://doi.org/10.1016/j.tra.2019.09.054.

Bwambale, A., Choudhury, C.F., Hess, S., Iqbal, MdS., 2020. Getting the Best of Both Worlds: a Framework for Combining Disaggregate Travel Survey Data and Aggregate Mobile Phone Data for Trip Generation Modelling. https://doi.org/10.1007/s11116-020-10129-5. Transportation.

Calabrese, F., Di Lorenzo, G., Liu, L., Ratti, C., 2011. Estimating Origin-Destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area. IEEE Pervasive Comput. 10, 36–44.

Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M., 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. Transport. Res. C Emerg. Technol. 68, 285–299.

[4] GSM/WCDMA/LTE/5G.

Diao, M., Zhu, Y., Ferreira, J., Ratti, C., 2016. Inferring individual daily activities from mobile phone traces: a Boston example. Environ. Plan. B Plan. Des. 43, 920–940. https://doi.org/10.1177/0265813515600896.

Dypvik Landmark, A., Arnesen, P., Södersten, C.-J., Hjelkrem, O.A., 2021. Mobile phone data in transportation research: methods for benchmarking against other data sources. Transportation 48, 2883–2905. https://doi.org/10.1007/s11116-020-10151-7.

Edwards, A.W.F., 1992. Likelihood. CUP Archive.

Guerrero, T.E., Guevara, C.A., Cherchi, E., Ortúzar, J. de D., 2020. Addressing endogeneity in strategic urban mode choice models. Transportation. https://doi.org/10.1007/s11116-020-10122-y.

Gundlegård, D., 2018. Transport analytics based on cellular network signalling data. Linköping Studies in Science and Technology. Linköping University Electronic Press.

Gundlegård, D., Rydergren, C., Breyer, N., Rajna, B., 2016. Travel demand estimation and network assignment based on cellular network data. Comput. Commun. 95, 29–42.

Heckman, J.J., 1978. Dummy endogenous variables in a simultaneous equation system. Econometrica 46, 931–959. https://doi.org/10.2307/1909757.

Houston, D., Luong, T.T., Boarnet, M.G., 2014. Tracking daily travel; assessing discrepancies between GPS-derived and self-reported travel patterns. Transport. Res. C Emerg. Technol. 48, 97–108.

Indebetou, Lovisa, Alexander, Börefelt, 2018. Resvanor Invånare 30-49 År I Umeå Tätort -Kartläggning Med Hjälp Av Ny Datainsamlingsmetod Hösten 2017. Umeå.

Internetstiftelsen, 2019. Svenskarna och internet.

Janzen, M., 2019. Simulating Annual Long-Distance Travel Demand (PhD Thesis). ETH Zurich.

Kristoffersson, I., Daly, A., Algers, S., Svalgård-Jarcem, S., 2020. Representing travel cost variation in large-scale models of long-distance passenger transport (No. 2020:6). Working Papers in TransportEconomics.

Manski, C.F., Lerman, S.R., 1977. The estimation of choice probabilities from choice based samples. Econ. J. Econ. Soc. 1977–1988.

McFadden, D., 1974. In: Zarembka, P. (Ed.), Analysis of Qualitative Choice Behavior, Frontiers in Econometrics. Academic Press, New York, NY.

Prelipcean, A.C., Susilo, Y.O., Gidófalvi, G., 2018. Collecting travel diaries: current state of the art, best practices, and future research directions. Transp. Res. Procedia 32, 155–166.

Stopher, P.R., Greaves, S.P., 2007. Household travel surveys: where are we going? Transport. Res. Part Policy Pract. 41, 367–381.

Stopher, P., FitzGerald, C., Xu, M., 2007. Assessing the accuracy of the sydney household travel survey with GPS. Transportation 34, 723–741.

Tolouei, R., Psarras, S., Prince, R., 2017. Origin-destination trip matrix development: conventional methods versus mobile phone data. Transp. Res. Procedia 26, 39–52.

Toole, J.L., Colak, S., Sturt, B., Alexander, L.P., Evsukoff, A., González, M.C., 2015. The path most traveled: travel demand estimation using big data resources. Transport. Res. C Emerg. Technol. 58, 162–177.

Varela, J.M.L., Börjesson, M., Daly, A., 2018. Quantifying errors in travel time and cost by latent variables. Transp. Res. Part B Methodol. 117, 520–541. https://doi.org/10.1016/j.trb.2018.09.010.

Walker, J., Li, J., Srinivasan, S., Bolduc, D., 2010. Travel demand models in the developing world: correcting for measurement errors. Transport. Lett. 2, 231–243. https://doi.org/10.3328/TL.2010.02.04.231-243.

Wu, X., Guo, J., Xian, K., Zhou, X., 2018. Hierarchical travel demand estimation using multiple data sources: a forward and backward propagation algorithmic framework on a layered computational graph. Transport. Res. C Emerg. Technol. 96, 321–346. https://doi.org/10.1016/j.trc.2018.09.021.