Mode choice latent class estimation on mobile network data

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

a VTI Swedish National Road and Transport Research Institute
Malvinas väg 6
114 28 Stockholm
SWEDEN

b Communications and Transport Systems, Department of Science and Technology (ITN), Linköping University
Kommunikation och transportsystem, ITN
601 74 Norrköping
SWEDEN

c Economics, Department of Management and Engineering (IEI), Linköping University
Institutionen för ekonomisk och industriell utveckling
Avdelningen för nationalekonomi
581 83 Linköping
SWEDEN

d ITS, University of Leeds
34-40 University Road
Leeds LS2 9JT
UNITED KINGDOM

Declaration of interest: There is no conflict of interest.

Corresponding author: Angelica Andersson angelica.andersson@vti.se

Mode choice latent class estimation on mobile network data

Abstract
In this paper we use a nested latent class logit specification to define and estimate a large-scale mode choice demand forecasting model. We estimate this model based on mobile phone network data translated to roughly 100 000 long-distance trips within Sweden, achieving convergence of the model and credible parameter estimates. We develop methods to address two problems stemming from the nature of this data: the difficulties of distinguishing bus trips from car trips (since they share the same infrastructure) and distinguishing business from private trips (since trip purpose is unknown). To address the first issue, we estimate a nested logit model with an artificial nest that accounts for the differences in utility between bus and car. To address the latter issue, we estimate a latent class model, identifying classes of trips interpreted as private and business trips. Addressing these two issues substantially improves model fit.

Keywords: Demand model, mode choice, latent class, mobile phone network data, travel behaviour, long-distance travel

1 INTRODUCTION

Transport forecasting models are a cornerstone of transport appraisal. Accurate transport demand models supporting policy advice are urgently needed, given that the transportation sector, which contributes to roughly a third of all greenhouse gas (GHG) emissions in the EU, has increased emissions during the last decade, making it difficult to meet the EU's share of the 2-degree target in the Paris agreement (European Environment Agency, 2021).

In this paper we define and estimate a large-scale mode choice demand forecasting model for long-distance passenger trips based on mobile phone network data. The raw mobile network data consist of signals to or from antennae that the phones have connected to. To build a data set that can be applied in our mode choice estimation, we first extract trips from the raw data, identifying origin, destination and mode. Then traffic supply data for all modes are matched to this trip data.

Two key challenges in using this data in the mode choice model estimation arise from the nature of the resulting trip data. The first challenge is to distinguish bus trips from car trips. This cannot be determined with certainty in mobile network data because the identification of the choice outcome is based on the transportation infrastructure used, recognised from the antennae a phone has connected to during a trip. The second challenge is that information concerning trip purpose is not revealed by the

data.

A better understanding of how to use mobile phone network data and address such problems in travel demand modelling, which is the main aim of this paper, is essential because the traditionally used national travel surveys (NTS) suffer from low and declining response rates (De Heer and De Leeuw, 2002; Prelipcean et al., 2018). Hence, the surveys are becoming less representative of the full population over time. There is a risk that travellers with a high value of time are less likely to respond to the survey (Stopher and Greaves, 2007). If some modes are more prone to be underreported than others, this can greatly affect the resulting mode shares. Mobile phone network data has no nonresponse bias since it is collected passively, i.e. without active participation of the traveller. Janzen et al. (2018) note that in their French data, long-distance trips were underreported in their survey data compared to their mobile phone network data, which were about twice as frequent. In recent years GPS tracking surveys have been used to reduce response errors and retrieve more detailed route information. Unfortunately, GPS tracking surveys suffer from even lower response rates than traditional travel surveys (Indebetou and Börefelt, 2018).

A major advantage with mobile phone network data is that the operator can observe large numbers of trips at low cost. However, since the phone owners do not give explicit permission to use the data, it is necessary to clean the data of any personal information before it can leave the operators' servers (due to privacy laws such as GDPR). Moreover, links between trips made by the same phone owner are not allowed in the extracted data, since this could be used to identify the phone owner. Hence, identification of trips (including classification of main mode and a few other characteristics of the trips, such as time of day) from antennae signals was carried out on the operator's servers. The data that could leave the operator's servers was a random sample of such identified trips.

In this paper we estimate a mode choice forecasting model on such a data sample, including roughly 100 000 long-distance domestic trips (trip length >100 km) in Sweden. We estimate a nested logit latent class mode choice specification on this data, addressing two key problems related to mobile phone network data: the difficulty of distinguishing bus from car trips (sharing the same infrastructure) and that trip purpose is unknown to the modeller.

The first challenge, distinguishing bus trips from car trips, stems from the fact that there is no geospatial difference between bus and car in the original antenna data since both modes use the road infrastructure[1]. We extracted the trip data used in the estimation of the mode choice demand model by applying the mode identification method referred to as the Route/Antenna method, described in Breyer et al. (2021). The Route/Antenna method was applied at the operators' servers and could therefore use disaggregated data about antenna connections. In this method, the trip is classified as one of the modes rail, air, or road, based on antenna connections and infrastructure locations. In this paper, we address the first challenge by empirically evaluating three different methods of separating bus from car in the mode choice estimation model.

The second challenge relates to the difficulty of identifying business trips from private trips, since the data lacks information about the trip purpose, as well as socio-economic information about the traveller. Not being able to identify business and private trips is a particular problem since it is widely acknowledged that the mode choice and the value of travel time differ between business and private trips. In this paper, we address the second challenge by proposing and empirically evaluating three different ways of incorporating indications of trip purpose derived from mobile network data into the model. The first two are based on a business trip indicator, and the third is a latent class model with two classes that we interpret as private and business trips.

[1] The mobile phone data also likely includes some observations from truck drivers on the roads. A crude calculation based on official vehicle kilometre statistics implies that the share of truck drivers among road-travellers undertaking long-distance trips is around 5 percent. Unlike for bus, we have no supply data available for trucks and therefore have no possibility to separate truck observations from car observations.

Most research applying mobile phone network data in the analysis of passenger transport has so far focussed on the extraction of origin-destination-matrices (Alexander et al., 2015; Bekhor et al., 2013; Caceres et al., 2020, 2013; Calabrese et al., 2011; Gariazzo et al., 2019; Gundlegård et al., 2016; Iqbal et al., 2014; Tolouei et al., 2017; Toole et al., 2015). Dypvik Landmark et al. (2021) investigated the robustness and quality of mobile phone network data compared to other data sources and found that mobile phone network data is particularly useful for long-distance trips, due to the coarse spatial resolution. Mobile phone network data have been used on its own for route choice modelling (Bwambale et al., 2019b, 2019a), and also in combination with survey data for route choice purposes (Bwambale et al., 2020). Furthermore, Bwambale et al. (2019c) combined mobile phone data with GPS data to model departure time choice. Ghasri et al. (2017) use mobile phone data to replicate trip patterns using a decision tree classifier. Two authors have combined mobile phone network data with survey data on the OD level to estimate travel demand models. Janzen (2019) estimated an activity-based model based on a synthetic population, while Brederode et al. (2019) used a multi-proportional gravity model to fuse mobile phone network data with survey data on the OD level before parameter estimation. Burgdorf et al., (2020) validated the behavioural output of an off-the-shelf mobile phone network data provider with the behavioural output from a gravity model based on survey data, and found that the mode split and travel frequencies were similar. Huang et al. (2019) reviewed studies classifying modes of trips identified in mobile phone network data. Out of the 22 studies found by Huang et al. (2019) only four studies separate bus trips from car trips, where Danafar et al. (2017), Kalatian and Shafahi (2016) and Phithakkitnukoon et al. (2017) mainly focus on short distance trips based either on proximity to route or travel speed, and Wang et al. (2010) only consider an example origin-destination pair (OD pair) where there is a clear difference in travel time between car and mass transit. Yang et al. (2022) confirm the difficulty of distinguishing bus and car trips based on geospatial information, even in the case of location-based services data (which derived from a combination of sensors such as Wi-Fi, Bluetooth, cellular tower, and GPS information whenever a mobile application updates the phone's location). In their study bus trips had the least prediction accuracy of all considered modes, probably due to the similarity between bus and car trips.

To our knowledge, no other authors have estimated country-wide large-scale mode choice models on mobile network data, which can be used for forecasting. We propose a model structure addressing the challenge of distinguishing road modes in mobile phone data, as well as a model structure addressing the challenge of lack of trip purpose in mobile phone data. While Andersson et al. (2022) define a mode choice model estimated on mobile network data and conduct a first test estimation of the model, that model included fewer variables and did not address the two challenges of this paper. Our main contribution in this paper is to increase the understanding of how to use mobile phone data to estimate large scale transport demand models.

2 DATA

2.1 Mobile phone network data

The mobile phone network data utilised in this paper is based on antenna connections in Sweden from one mobile operator during one week in 2018 (pre-pandemic). The data contains billing data and location updates extracted from the core network and includes periodic, location area (LA), routing area (RA), tracking area (TA), and cell updates following the terminology proposed by Gundlegård (2018). The data consists of 936 million connection events from 2.7 million users.

To extract trips from the raw data (antenna connections), we process it in two steps, using a remote-access setup (as in de Montjoye et al., (2018)). Since the phone owners did not give explicit permission to use the data, the code is brought to the raw data on the operator's servers (due to privacy laws). The first step is to extract trips from antennae observations and assign each trip to a start zone and a destination zone. We use a stop-based trip extraction method as described in (Breyer et al., 2021), where a stop is assumed when the user does not move more than two kilometres for more than

two hours. We apply the same zones as the Swedish long-distance transport model (owned by the Transport Administration). In total, there are 682 such zones in Sweden.

The second step is to classify each trip in the sample by travel mode. We classify the main mode of a trip as air if the trip contains two events that are within 10 km distance to an airport and where the distance between the two events is at least 200 kilometres and straight-line travel speed at least 200 km/h. For the remaining trips, we use the geometric Route/Antenna mode classification method described in Breyer et al. (2021) to identify the main mode of each trip. The method identifies the mode of each trip by comparing the spatial distribution of antennae that are connected during the trip to the fastest car and rail routes (according to OpenTripPlanner (2020)) from the origin to the destination. See Breyer et al. (2021) for a more thorough description of the mode identification process.

The resulting anonymised and mode-classified trips include no personal information and can be exported from the operator's servers. Moreover, links between trips made by the same phone owner are not allowed in the data extracted from the operator's servers, since this could be used to identify the phone owner. We select a random sample of 100 000 long-distance trips from this data, to estimate the mode choice demand model outside the operators' servers. Long-distance trips are defined as trips with a distance of at least 100 kilometres between origin and destination zone centroids. The variables in this trip data are described in Table 1.

***Table 1**: Features in the mobile phone network data*

Feature	**Values**
Origin	Origin zone of the trip
Destination	Destination zone of the trip
Peak hour (r)	True if trip start is during peak hours (Mon-Fri at 7-9 or 15-18), false otherwise
Weekend (s)	True if day is Saturday or Sunday, false otherwise
Business Departure Time	True if departure time is between Monday 12 am and Friday 12 am, false otherwise
Employment	True if the user that made the trip has a regular day activity (see Appendix A), false otherwise
Daytrip (d)	True if the user has made a return trip the same day, false otherwise
Regular home/night location (h)	True if the user that made the trip has a regular night-time location (see Appendix A), false otherwise
Identified mode (m)	Most likely main mode of the trip

There are some deficiencies in the dataset. First, there is a risk of errors in the mode identification performed using the methods in Breyer et al. (2021). One reason for this is the assumption that all road travellers use the fastest route (according to Open Trip Planner), while there are in fact several routes that are used for many origin-destination pairs in Sweden. The fastest route may also vary among departure times. For rail, the number of used routes for a given OD pair is rarely larger than one. Hence, there is a risk that some road trips (not using the fastest route) are identified as rail trips (but the reverse is less likely). A more thorough analysis of what type of situations are at higher risk of being misclassified can be found in Andersson et al. (2022). Validation on a smaller dataset between Norrköping and Linköping revealed that the Route/Antenna method had an accuracy of 95.5%. That is, when using 510 trips with manually annotated modes of transport, 95.5% of the observations had a match between the most probable mode from the Route/Antenna method and the annotated mode (Breyer et al., 2021). However, for the particular origin-destination pair Norrköping-Linköping it is uncommon to use another route than the fastest (straight on the motorway E4 connecting the two cities).

Second, data includes no trip purposes. Instead, we construct a business trip indicator B from an

indicator of whether the traveller is employed ("Employment"[2] in Table 1) combined with an indicator of whether the trip started during business hours ("Business Departure Time" in Table 1). If the trip is performed by a traveller who is labelled as employed, and the long-distance trip started during business time, we take the business trip indicator B to be one, and zero otherwise.

2.2 Supply data

Transport supply data for the chosen and unchosen modes is also needed to estimate the model. Such data were provided by the Swedish Transport Administration. The Administration obtained this supply data by simulating travel time and cost matrices for all pairs of origin and destination zones using the Swedish long-distance transport model, using Emme4[3] for network assignment. The costs are at the price level of 2017, and we adjusted them to 2018 prices (which is the year the mobile phone network data were collected).

For public transport trips, the supply data includes a dummy variable for availability[4] (a), ticket price (c), in-vehicle travel time (t), half headway of main mode (w), and number of boardings by OD pair (n). The data also contains the sum of the road distance between the centroid of the start zone and closest mode terminal (airport, train station or bus terminal), and the distance between the centroid of the destination zone and closest mode terminal (δ). As the mode and price of the connecting trip is unknown, this sum of distances is included as a representation of the travel time and travel cost of the connecting trips.

For car trips, the supply data includes travel time (t) and travel distance by OD pair. It is well known that the marginal cost of car travel is subject to great variation between trips and vehicles (Kristoffersson et al., 2020), due to variation in energy efficiency, fuel type, driving style, traffic environment, and the type of car access (car ownership, private leasing or company car influence the marginal cost of car use, both through prices and taxes, and depreciation). We therefore approximate the marginal cost of car use. The Swedish tax authority allows the employer to reimburse employees using their private car for a business trip by maximum 1.85 SEK/km.[5] Roughly half of this amount accounts for fuel costs, whilst the rest covers depreciation which depends largely on the car. The average occupancy for private long-distance car trips is just over two persons per car (2.22) (Trafikverket, 2020), while we assume that it is just one person for business trips. We therefore approximate a car distance cost of $1.85/2 \approx 0.9$ SEK/km for private trips and 1.85 SEK/km for business trips.[6]

There are probably some substantial measurement errors in the supply data. Varela et al. (2018) show that the measurement error in the travel cost is larger than the measurement error in the travel time, not only for car trips. For air and rail trips, there are large variation in ticket prices depending on the traveller (discounts for children and retirees are common), the timing of the trip (departure time and day) and how long in advance the ticked is booked. Such variation likely causes a larger attenuation bias in the cost parameter than in the time parameter, implying that the value of time calculated from the estimated time and cost parameters would be overstated.

[2] The employment indicator is constructed based on trips made by the same traveller in one week, including also regional trips (see Appendix A).
[3] https://www.inrosoftware.com/en/products/emme/
[4] A mode is set to unavailable if the origin and destination zones share the same closest terminal (airport, train station or bus terminal) of that mode.
[5] Inkomstskattelagen (income tax law) 12 kap. 5 §, 2007.
[6] 10 SEK is approximately €1.

2.3 Descriptive statistics

As shown in Table 3, 26% of travellers in the dataset are identified as employed, and 55% of trips are performed during business time according to the definition in Table 1. This leads to 13% of trips being labelled as business trips. Moreover, 26% of trips are day trips (meaning the return trip is performed the same day), 79% of travellers are identified as having a regular night location ("Regular home/night location", in Table 1). Finally, 23% of the trips started during peak hours (as defined in Table 1), and 34% of trips started during the weekend.

Of the long-distance trips in the sample, 1.6% are classified as air trips, 32.2% as rail trips, and 59.9% as road trips. The remaining 6.3% could not be classified by the Route/Antenna method as no appropriate routes have been found. These trips are likely to be road trips not using the fastest route (see Section 2.1). These non-classified trips are excluded from the estimation described in Section 3.

Next, we compare these mode shares to the Swedish national travel survey (NTS). The most recent NTS was conducted in 2020 and 2019 (Trafikanalys, 2021a). However, in this survey respondents were only asked to report trips that they made during one pre-determined survey day. Since long-distance trips are rare, there are few observations of long-distance trips in this data, making statistical analysis unreliable.

In previous surveys, conducted 2005-2006 and 2011-2016, respondents were asked to report any long-distance trips performed during the past few months[7], which provided more observations of such trips. In the survey conducted 2011-2016, the reported domestic long-distance mode shares are 15.2% air, 14.3% rail, 67.5% road (car + bus), and 3% other (see Table 2). The 15.2% share of air trips is substantially larger than the corresponding share of 2% in the data extracted from mobile network data. However, the surveys conducted 2011-2016 have low response rates with an average of 41% (Holmström, 2017; Holmström and Wiklund, 2015), as well as a higher response rate among older respondents. A major reason for this is that respondents were recruited by telephone and that many households (in particular older) at that time still had a landline telephone connection or a registered telephone number (which is no longer the case). However, younger households had fewer landline phones, and were thus more difficult to recruit. As many as 48 percent of the respondents (a random sample of the population) could not be reached, because the survey firm had no telephone number for them, or the respondent did not pick up the phone. In the preceding 2005-2006 NTS, only 12 percent of the respondents could not be reached: most households still had a landline phone at that time (Trafikanalys, 2018). For this reason, the 2005-2006 is probably more reliable and representative of the full population than the subsequent NTS data, and therefore we compare the modal splits in our mobile phone network data with the mode shares of the 2005-2006 NTS data.

To make the 2005-2006 NTS data as representative of 2018 as possible (the year of mobile phone network data collection), we adjust the resulting modal split according to the change in aggregate numbers of trips by mode in Sweden. We use national statistics on the change in the number of long-distance rail trips (Trafikanalys, 2011, 2018a) and domestic air trips (Trafikanalys, 2021b). For car trips we only had access to changes in the total number of vehicle kilometres driven (Trafikanalys, 2018b), but no data on changes in long-distance car travel in particular. We still used this data as a rough approximation of the change in the number of long-distance car trips.

Table 2 depicts the resulting mode shares for the various data sources. The mobile phone network data is in line with the more reliable survey data from 2005-2006, adjusted to 2018 as described above. However, the table indicates that the share of air trips is overstated in the 2011-2016 NTS data, confirming the suspicion that the low and biased response rate has implied that the trips in the 2011-

[7] In the 2011-2016 survey, respondents reported trips longer than 100 km during the past month and trips longer than 300 km during the last three months, and in the 2005-2006 survey respondents reported trips longer than 100 km during the last two months.

2016 data are not representative of the full population.

***Table 2**: Domestic Swedish mode shares from mobile phone network data, national travel survey data 2005-2006, national travel survey data 2011-2016 and 2005-2006 adjusted to 2018.*

	Rail	***Air***	***Road (car/bus)***	***Other/unclassified***
Mobile phone data 2018	*32%*	*2%*	*60%*	*6%*
NTS 2005-2006	*12%*	*4%*	*82%*	*2%*
NTS 2011-2016	*14%*	*15%*	*68%*	*3%*
NTS 2005-2006 adjusted to 2018	*15%*	*4%*	*79%*	*2%*

The greatest difference between the mobile phone network data and the adjusted 2005-2006 NTS data is that the share of rail trips is considerably larger in the former. A possible explanation is that some of the car trips were incorrectly classified as rail trips by the Route/Antenna method used to identify the mode of each trip in the mobile phone network data. As discussed previously, this is likely to happen for OD pairs where car trips do not always take the fastest route (or when the fastest route may vary between departure times). Another possible reason is that rail trips are underreported in the 2005-2006 NTS data due to fatigue or forgetfulness. This is still not likely, since car trips, rather than rail trips, tend to be underreported in NTS data for this reason (WSP Analysis and Strategy, 2012).

Availability rates for rail, bus, air and car for the observed trips are 99.6%, 98.3%, 77.0% and 100% respectively (meaning the average availability of modes for each of the four modelled modes for all observed OD pairs). Supply statistics for all alternative travel modes coded for all observed trips, as well as the trip specific information from the mobile phone network data for all observed trips are presented in Table 3.

Table 3: Statistics of model variables.

Variable:	**Minimum:**	**Maximum:**	**Mean:**	**Median:**	**Standard deviation:**
Business departure time [t/f]	0.0	1.0	0.55	1.0	0.5
Daytrip [t/f]	0.0	1.0	0.26	0.0	0.44
Regular home/night location [t/f]	0.0	1.0	0.79	1.0	0.41
Employed [t/f]	0.0	1.0	0.26	0.0	0.44
Car travel time [min]	56.4	1136	185	152	102
Car cost private [SEK]	90.1	1686	265	214	152
Car cost business [SEK]	185	3465	544	440	311
Bus travel time [min]	35.0	1790	324	277	170
Bus number of boardings	1.0	7.9	2.89	3.0	1.19
Bus first wait time [min]	2.55	480	76.7	53.3	82.7
Bus travel price [SEK]	74.3	1059	196	162	102
Air travel time [min]	21.6	302	118	129	42.8
Distance to airport [km]	6.23	383	92.9	81.6	47.9
Air first wait time [min]	12.6	480	69.5	48.0	68.4
Air number of boardings	1.0	3.0	1.69	2.0	0.54
Air travel price [SEK]	417	4582	2035	2097	890
Rail travel time private [min]	11.1	1317	192	170	109
Distance to train station private	0.34	503	22.1	12.8	28.3

[km]					
Rail first wait time private [min]	2.82	240	35.4	28.3	28.2
Rail number of boardings private	1.0	5.48	1.95	2.0	0.77
Rail travel price [SEK]	5.7	2017	749	666	364
Rail travel time business [min]	11.1	1200	190	169	105
Distance to train station business [km]	0.34	503	22.0	12.7	28.3
Rail first wait time business [min]	2.82	240	32.6	28.2	23.9
Rail number of boardings business	1.0	6.32	2.07	2.0	0.84
Business dummy	0	1	0.13	0.0	0.34
Weekday peak hour	0	1	0.25	0.0	0.43
Weekday off-peak	0	1	0.46	0.0	0.50
Weekend	0	1	0.29	0.0	0.45

3 MODEL SPECIFICATIONS

In this section we specify a series of logit model specifications. We apply maximum likelihood (MLE) (Edwards, 1972) in the estimation. A short summary of all specifications can be found in Appendix C.

3.1 Modelling the road modes

Since the modes bus and car are not distinguished in the data, we formulate three different model specifications I-III, modelling the bus and car alternatives in different ways, described below.

3.1.1 Joint utility function for road choice

Specification I is a multinomial logit (MNL) model, in which car and bus are modelled as one and the same mode. The utility function for this combined road mode is a weighted average of the utility functions for bus and car, where the weights equal the mode shares for long-distance bus (π) and car trips ($1-\pi$) in Sweden

$$V_{road} = ASC_{road} + \beta_{t\_road}(t_{car}(1-\pi) + t_{bus}\pi) + \beta_c(c_{car}(1-\pi) + c_{bus}\pi) + \beta_a a + \log(w_{bus})) + \beta_s s + \beta_r r, \quad (1)$$

where t denotes travel time (in-vehicle time only), c denotes travel cost, a is a dummy taking the value 1 if there is a scheduled bus connection available between origin and destination zones and 0 otherwise, s is set to 1 if the trip starts during a weekend, 0 otherwise. The dummy variable r takes the value 1 if the trip starts during peak hours, 0 otherwise, n is the number of boardings during the bus trip and the waiting time w is taken to be half the headway of the main mode. The alternative specific constant ASC_{road} has been fixed to 0. The share π is observed in the national travel survey from 2011-2016 (implying $\pi = 6\%$).

For air trips, the utility function is specified as

$$V_{air} = ASC_{air} + \beta_{t\_air}t_{air} + \beta_c c_{air} + \beta_{\delta\_air}\delta_{air} + \beta_n n_{air} + \beta_w \log(w_{air}), \quad (2)$$

where ASC_{air} is the alternative specific constant of air, t_{air} denotes the in-vehicle travel time for air, c_{air} is the ticket price of the air trip, δ_{air} is the distance between the zone centroid and the closest airport (representing access and egress trips), n_{air} is the number of boardings needed for the air trip

and w_{air} is taken to be half the headway of the main mode.

For rail trips, the utility function is defined as

$$V_{rail} = ASC_{rail} + \beta_{t\_rail} t_{rail} + \beta_c c_{rail} + \beta_{\delta\_rail} \delta_{rail} + \beta_n n_{rail} + \beta_w \log(w_{rail}), \quad (3)$$

where ASC_{rail} is the alternative specific constant of rail, t_{rail} denotes the in-vehicle travel time for rail, c_{rail} is the ticket price of the rail trip, δ_{rail} is the distance between the zone centroid and the closest train station (representing access and egress trips), n_{rail} is the number of boardings needed for the rail trip and w_{rail} is half headway of the main mode.

3.1.2 Random assignment to bus or car

Specification II is identical to specification I, except that the two road modes bus and car are specified as separate alternatives. All road trips in OD pairs without any bus connection are assumed to have been made by car. All other road trips are then randomly assigned either to bus or car, such that the share of bus and car trips among all OD pairs equals the mode shares π and $1 - \pi$, respectively. This specification is similar to the approach proposed by Beser Hugosson (2003).

The utility function for car is given by

$$V_{car} = \beta_{t\_car} t_{car} + \beta_c c_{car} + \beta_s s + \beta_r r. \quad (4)$$

The utility function for bus is given by

$$V_{bus} = ASC_{bus} + \beta_{t\_bus} t_{bus} + \beta_c c_{bus} + \beta_n n_{bus} + \beta_w \log(w_{bus}). \quad (5)$$

3.1.3 Composite utility

Specification III is identical to specification II, except that the bus and car alternatives are assumed to belong to a separate nest, i.e., specification III is a nested logit model. This implementation of the bus and car alternative in a joint nest resembles the approach used in Daly et al. (2002). Here, the probability of choosing mode $i \in$ {car, bus} within the road nest is given by

$$p_{road,i} = Pr(road) Pr(i \mid road) = \frac{\exp V^*_{road} \quad \exp V_i}{\sum_k \exp V^*_k \sum_j \exp V_j} \quad (6)$$

where $V^*_{road} = \theta \log \sum_j \exp(V_j/\theta)$, and V_j is the utility for mode $j \in$ {car, bus} given by Equations (4) and (5) (V_{bus} is set to $-\infty$ when there is no available bus connection), and $k \in$ {road, rail, air}. 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. θ is estimated along with the other parameters but restricted to the bounds $0 < \theta \leq 1$. The log-sum V^*_{road} is the maximum expected utility and can be interpreted as a composite utility function for car and bus, where the balance between the modes depends on the generalised costs of the modes. This specification makes it possible to connect the choice to one of the three utilities V_{road}, V_{air} or V_{rail}, while still estimating parameters for the four modes car, bus, air, and rail.

Since none of the specifications I-III are nested versions of each other, the likelihood ratio test cannot be used to compare the relative performance of the specifications. Instead, we use the Akaike

information criterion (AIC). The AIC measure rewards goodness of fit and gives a penalty for the number of parameters in the model. The penalty indirectly discourages overfitting, as a high number of explanatory variables increases the risk of overfitting. The AIC is calculated as

$$AIC = 2k - 2LL \tag{7}$$

where k is the number of estimated parameters and LL is the log likelihood.

It is not fair to directly compare the AIC of specification II with specifications I or III since the LL of specification II is calculated based on four alternatives (as the road alternative is randomly assigned to either bus or car before estimation). This contrasts to specifications I and III in which the LL is calculated based on the three original alternatives in the choice set. To compensate for this one could add an additional likelihood for specifications I and III, representing the choice for road travellers between car and bus, calculated as

$$l_{road,a}(p_{bus|road}\log(p_{bus|road}) + p_{car|road}\log(p_{car|road})), \tag{8}$$

where $l_{road,a}$ is the number of trips labelled as road where both bus and car are available options, $p_{bus|road}$ is the share of bus trips among the road trips and $p_{car|road}$ is the share of car trips among the road trips (see Appendix B for details of the derivation of this expression). The expression in Equation (8) would then reflect the additional burden of making a random assignment to either bus or car for all the choices in which road was chosen and both bus and car are available.

3.2 Trip purpose

Specification III will prove to be superior to specifications I and II (see Section 4). For that reason, subsequent specifications IV-VII implementing trip purpose in different ways will incorporate the implementation of the two road modes as in specification III.

Specifications IV-VI use the business trip indicator B in different ways, where specification VI is a latent class model (Hess, 2014). In specification VII, which is also a latent class model, we use more variables to indicate whether a trip is a business trip.

3.2.1 Business trip indicator specifications

In specification IV, the business trip indicator B is simply added as a dummy variable in the utility function for car (Equation (4)). Specification V is segmented based on the business trip indicator, i.e., different models are estimated for trips identified as business (B=1) and private (B=0). For the business trips, the weekend variable in the utility function for car (Equation (4)) is removed since by definition no business trips start during the weekends (see Section 2). Moreover, the bus alternative is discarded from the model for business trips; according to the NTS data, very few long-distance business travellers use bus.

3.2.2 Two latent class specifications

In specifications VI and VII we use the data indicating the trip purpose (business trip or private trip) but take into account that it is not fully observed. We assume that the trips can be classified into the two latent classes: business trips and private trips. The two unconditional class probabilities, q_p and q_b , where the subscript p denotes private and b denotes business, are determined by the multinomial logit specification. The probability that the trip is private is $q_p = 1 - q_b$.

The conditional probability that mode m is observed for a trip, given the parameter vectors β, is

$$P(m|\beta) = q_p P_p(m|\beta_p) + q_b P_b(m|\beta_b) \tag{9}$$

$P_p(m|\beta_p)$ and $P_b(m|\beta_b)$ represent the probability that mode m is chosen conditional on the trip purpose parameters for business and private, respectively. The utility functions by mode for business and private trips include a similar set of variables, but different parameters β_p and β_b are estimated. The bus mode is omitted from the business model.

In model specification VI the probability that a trip's purpose is business is only indicated by the business trip indicator and is given by

$$q_b = \frac{1}{1 + e^{-(\gamma + B\gamma_B)}}$$

where B is the business trip indicator, γ and γ_B are parameters to be estimated. Specification VI collapses to specification V if $\gamma = -0.5\gamma_B$ and $\gamma_B \to \infty$ and thus can be seen as a generalised version of specification V, which means that the likelihood ratio test can be used to compare the specifications.

Our final specification VII models the class probabilities in a more refined way, namely $q_b = \frac{1}{1 + e^{-(\gamma + d\gamma_d + h\gamma_h)}}$ where γ, γ_d and γ_h are parameters to be estimated (relating to a constant, the day trip and regular home/night location variables respectively defined in Table 1). Specification V or specification VI is not a restricted version of specification VII, so we cannot compare these specifications using the likelihood ratio test. Instead, we use the AIC value as an indicator of the preferred specification.

The utility functions of specification VII are similar to those of specification VI, but in specification VII the parameters for the number of transfers for private trips, private air in-vehicle time and private distance to airport have been excluded due to insignificance (air travel time was found to correlate with the alternative specific constant in the private component of specification VI).

4 RESULTS

This section presents the estimation results. All estimations were performed using Biogeme version 3.2.5 (Bierlaire, 2020). The chosen optimisation algorithm was the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm.

4.1 Identifying the road mode

The estimation results for the three specifications I-III, differing in the way that the two road modes are specified, are shown in Table 4. As explained in Section 2, we cannot link trips made by the same traveller, and are thus not able to consider correlation between repeated observations. For this reason, we report the robust t-values given by Biogeme (Bierlaire, 2020), allowing for some specification error, including a possible violation of the assumption that the observations are independent.

In the nested specification III, θ reaches its upper boundary 1, implying that the nested multinominal model collapses to an MNL. This indicates that the modes within the road nest are not more similar to each other than to the modes. Specification III still differs from the other two models.

From the AIC test presented in Table 4, we conclude that specification III has the lowest AIC value of the three and thus is the preferred specification. Hence, the composite utility using a road nest fits the data best, and is thus the preferred method to apply to mobile network data, in which bus and car cannot be distinguished.

***Table 4**: Estimation results for handling the road mode.*
**Structural coefficient θ restricted to its maximum 1 to maintain consistency with intuitive response characteristics.*

	Specification I		Specification II		Specification III	
Name	**Parameter value**	**Robust t-test**	**Parameter value**	**Robust t-test**	**Parameter value**	**Robust t-test**
ASC_{bus}			-2.39	-37.5	4.04	9.43
ASC_{air}	-1.04	-8.23	-0.75	-6.44	-0.817	-6.66
ASC_{rail}	-0.0891	-1.44	0.15	3.74	0.344	7.71
β_a	0.505	7.9				
β_c	-0.00065	-14.2	-0.00062	-14.1	-0.000584	-13.7
β_{δ_air}	-0.0197	-24.2	-0.0194	-24	-0.0192	-23.8
β_{δ_rail}	-0.0171	-47.7	-0.0157	-45.5	-0.0158	-44.9
β_w	-0.0186	-2.75	-0.128	-13.8	-0.14	-13.3
β_n	-0.0711	-10.6	-0.165	-15.4	-0.26	-21.6
β_r	-0.231	-13.3	-0.219	-12.7	-0.23	-12.7
β_{t_bus}			-0.00401	-28	-0.0361	-10.3
β_{t_car}			-0.00764	-58.8	-0.00771	-54.2
β_{t_air}	-0.0169	-15.8	-0.0143	-14.4	-0.0123	-12.9
β_{t_road}	-0.00794	-60.3				
β_{t_rail}	-0.00634	-45.3	-0.00527	-37.9	-0.00514	-34.3
β_s	0.37	21.4	0.349	20.5	0.377	20.8
θ					1	*
# parameters:	13		14		15	
Log likelihood:	-60788.43		-70199.6		-60485.61	
LL adjustment:	-9440.78		0		-9440.78	
Adjusted LL:	-70229.21		-70199.6		-69926.39	
AIC:	140484.4		140427.2		139880.8	
Number of observations:	92011		92011		92011	

4.2 Trip purpose

In the previous section specification III proved to be superior to specifications I and II. For that reason, subsequent specifications IV-VII incorporate the implementation of the two road modes as in specification III.

4.2.1 Business trip indicator

The estimated parameters for specification IV, the unsegmented model including the business trip indicator β_B, are shown in the first column of Table 5. In this case, the business trip indicator proved to be insignificant. In the subsequent columns, separate models are instead estimated for private and business trips in specification V.

To compare the segmented specification V with the unsegmented specification IV, a likelihood ratio χ^2 test is applied. The likelihood-ratio test statistic

$\lambda_{\mathrm{LR}} = 2(\mathrm{LL}_{V\_p} + LL_{V\_b}) - 2LL_{IV}$=148.588,

is χ^2-distributed. The segmented specification V has 25 degrees of freedom, and the unsegmented specification IV has 15 degrees of freedom. The value of the χ^2-distribution for 10 degrees of freedom and 0.001% significance level is 29.588. Hence the null hypothesis is rejected and the segmented model specification V is preferred.

4.2.2 Latent class model results

The latent class specification VI is presented in subsequent columns of Table 5. As mentioned in Section 3.2.2, we can use the likelihood-ratio test to compare specifications V and VI with the statistics

$\lambda_{\mathrm{LR}} = 2LL_{VI} - 2(LL_{V\_p} + LL_{V\_b})$ = 1367.272.

The segmented specification V has 25 degrees of freedom, and the latent class specification VI has 27 degrees of freedom. The value of the χ^2-distribution for 2 degrees of freedom and 0.001% significance level is 13.816. Hence the null hypothesis is rejected, and the latent class model specification VI is the preferred specification.

The two parameters that turned out to be useful in separating the classes private and business were γ_d (daytrip) and γ_h (regular home/night location). It is in line with expectations that the long-distance trip being a daytrip increases the probability that it is a business trip. Furthermore, there are groups of individuals that are less likely to have a regular home location during the nights (identified by the regular home/night location variable, see Appendix A). These groups include professions like police, security guards, nurses who travel at night (like ambulance nurses, or nurses taking care of sick elderly people in their homes), taxi or public transport drivers, as well as young adults spending the night out. As none of these groups are expected to make long-distance business trips very often, it is in line with expectations that the parameter γ_h is positive in explaining the probability of belonging to the business class.

The mean value of the business class probability q_b is 11%, which is close to the share of business trips from the 2011-2016 NTS of 12%. One could also consider investigating posterior class probabilities, which take information into account about each traveller to improve the probability of belonging to a certain class (for example, a traveller that has chosen to travel by air is more likely to be a business traveller) (Hess, 2014). However, in our dataset, very little is known about each traveller and even if the same person has made several of the long-distance trips this is not visible in the data. For this reason, posterior class probabilities are not used here.

Table 5: *Estimation results for evaluation of segmentation by trip purpose indicator. The structural coefficient θ has been fixed to 1 in the estimations below. Presented t-values are the robust version of t-values. * Shared parameter between business and private.*

IV		V				VI				VII			
Unsegmented		Private		Business		Private		Business		Private		Business	
Value	t-test	Value	t-test	Value	t-test	Value	t-test	Value	t-test	Value	t-test	Value	t-test
4.08	9.57	4.14	9.97			5.43	9.55			5.3	9.6		
-0.82	-6.67	-0.859	-6.27	-1.68	-5.3	-2.4	-0.453	-0.779	-3.17	-5.94	-10.4	-0.596	-1.64
0.341	7.55	0.314	6.43	0.564	4.8	0.653	3.7	2.37	13.2	0.864	6.34	2.45	10.9
-0.00917	-0.415					-0.00177*	-0.0504	-0.00177*	-0.0504				
-0.000584	-13.7	-0.00079	-10.9	-0.000876	-10.1	-0.00252	-12.5	-0.000235	-3.78	-0.00233	-7.78	-0.000165	-2.28
-0.0192	-23.8	-0.0194	-21	-0.0165	-8.1	0.0046	1.36	-0.0264	-17.4			-0.0286	-12.7
-0.0158	-44.9	-0.0155	-41.9	-0.0184	-17.5	-0.0229	-28	-0.0358	-15.6	-0.0222	-19.6	-0.039	-13.9
-0.14	-13.4	-0.139	-12.4	-0.142	-4.89	-0.131	-4.2	-0.463	-14.4	-0.138	-4.51	-0.478	-13.1
-0.26	-21.6	-0.249	-19	-0.258	-8.49	0.0711	2.03	-1.07	-15			-1.1	-11.5
-0.23	-12.6	-0.231	-11.3	-0.221	-5.56	-0.29	-4.68	-0.45	-8.55	-0.3	-5.2	-0.471	-8.71
-0.0365	-10.6	-0.0371	-11							-0.0439	-10.4		
-0.00771	-55.9	-0.00727	-45.2	-0.00967	-25.5	-0.0147	-16.8	-0.0141	-29.4	-0.0137	-10.6	-0.0146	-29.4
-0.0123	-12.9	-0.0142	-12.2	-0.00819	-4.11	-0.0911	-1.5	-0.0111	-6.41			-0.0128	-6.13
-0.00514	-35.9	-0.00468	-29.3	-0.00581	-12.6	-0.00195	-3.86	-0.0172	-19.2	-0.00235	-3.39	-0.0181	-19.7
0.375	20.2	0.372	19.8			0.763	14.5			0.72	11.7		
						0.187*	3.03	0.187*	3.03	-0.0166*	-0.161	-0.0166*	-0.161
										0.103*	2.99	0.103*	2.99
										0.131*	4.49	0.131*	4.49
15		14		11		27				25			
92011		79877		12134		92011				92011			
-60485.53		0		-8394.876		-59727.6				-59709.64			
121001.1		104060.7		16811.75		119509.2				119469.3			

	ASC_{bus}	ASC_{air}	ASC_{rail}	β_B or γ_B	β_c	β_{δ_air}	β_{δ_rail}	β_w	β_n	β_r	β_{t_bus}	β_{t_car}	β_{t_air}	β_{t_rail}	β_s	γ	γ_d	γ_h	**# parameters:**	**Sample size:**	**Log**	**AIC:**

The estimation result of the refined latent class specification VII is presented in the final column of Table 5. The AIC is lower for specification VII than for specification VI, and specification VII is thus the preferred specification.

The resulting values of travel time (VTT) for the preferred specification VII are presented in Table 6. As expected, the values of travel time are higher for business trips than for private trips. The VTT of business travel time are on the high side, but in line with the values of time derived from the current Swedish long-distance model. A standard assumption for business VTT is wage rate. A likely explanation is that *within the range of domestic business trips*, travel cost has a limited impact on the mode choice of long-distance business trips in Sweden. Travel time and scheduling constraints are likely much more important.

For private trips, the parameter for in-vehicle time was left out of the utility function for air trips since it was not significant. One reason could be that the travel time varies so little among domestic air trips. Another reason is that there are so few private air trips. This means that it is not possible to calculate the VTT for private air trips.

For private trips, bus has the highest VTT. A likely reason is that this is the most uncomfortable mode, with limited or no access to bathrooms and restaurants on board. However, VTT for bus may also not be credible for two reasons. First, the choice of long-distance domestic bus trips has been noted to be strongly influenced by party size in Swedish national travel surveys. As information about party size is of course not available in the mobile phone data, it was not possible to include it in the bus utility function. Second, a significant share of long-distance bus trips is sports teams or other organisations privately hiring a bus. For those trips, the generalized travel costs (including headway and fares) for scheduled buses that we apply are not representative.

Among business trips, air trips have the highest value of time, followed by rail trips and finally car trips. A higher VTT for rail than for car trips is not expected and indicates that travellers are not more productive (able to work more) on rail trips than while driving. Another explanation for this could be scheduling constraints, i.e., there may be a mismatch between scheduled meetings and available rail departures, especially for appointments early in the morning.

Table 6: *Values of travel time for different modes for the best-found model specification. The average currency exchange rate at the year of data collection was 1€ = 10SEK.*

	Value of travel time, private trips [€/h]	Value of travel time, business trips [€/h]
Air	-	465.0
Rail	6.1	659.4
Car	35.3	529.3
Bus	113.1	-

5 CONCLUSION

In this paper we show that mobile phone network data can be used for estimation of both state-of-

practice and advanced transportation mode choice models. We also show how to address two of the main challenges related to mobile phone network data. The first challenge is that bus trips and car trips are difficult to distinguish in mobile phone data since the mode identification is based on the proximity between network antennae and road infrastructure. We show that modes can successfully be distinguished for the purpose of forecasting models, by implementing a nested logit structure. The second challenge is that the trip purpose is unknown. We address this by identifying two groups of travellers with markedly different preferences in terms of the valuation of travel time, by estimating a latent class structure. These classes are interpreted as private and business travellers.

One weakness of mobile phone network data is the absence of socio-economic information about the traveller. For this reason, future studies could benefit from combining mobile phone network data with traditional data sources. Using the formulation of mobile phone network data presented in this paper, it would be reasonably straightforward to combine it with a traditional survey-based model. In this paper we have established the applicability of mobile phone network data for forecasting demand models using a traditional underlying model. However, the size of the dataset offered by mobile phone network data also opens up a range of alternative model formulations, among them: machine learning.

6 ACKNOWLEDGEMENT

This work was conducted within the Demopan project funded by the Swedish Transport Administration under Grant TRV 2018/126661. The authors would like to thank Clas Rydergren (Linköping University) and David Gundlegård (Linköping University) for fruitful discussions during the project time.

7 LIST OF ABBREVIATIONS

a	bus available
AIC	Akaike Information Criterion
ASC	Alternative Specific Constant
b	business (subscript)
B	business trip indicator
c	cost
c^*	chosen mode
δ	sum of the distance between the centroid of the start zone and closest mode terminal (airport, train station or bus terminal), and the distance between the centroid of the destination zone and closest mode terminal
d	daytrip
h	regular night/home location
i	observation
l	number of trips
m	mode
MNL	multinominal logit
n	number of boardings
na	bus not available
OD	origin-destination
p	private (subscript)
P	probability
q_x	probability that a trip's purpose belongs to class x (in latent class specifications)
r	peak hour
s	weekend
t	time
VTT	Value of Travel Time

w	first wait time (half headway)
β_x	utility function parameter of variable x
γ_x	class probability parameter of variable x
π	Share of bus trips in NTS 2011-2016
θ	Structural nesting parameter of specification III

8 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. Transp. Res. Part C Emerg. Technol., Big Data in Transportation and Traffic Engineering 58, 240–250. https://doi.org/10.1016/j.trc.2015.02.018

Andersson, A., Engelson, L., Börjesson, M., Daly, A., Kristoffersson, I., 2022. Long-distance mode choice model estimation using mobile phone network data. J. Choice Model. 42, 100337. https://doi.org/10.1016/j.jocm.2021.100337

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

Beser Hugosson, M., 2003. Issues in Estimation and Application of Long Distance Travel Demand Models. KTH Royal Institute of Technology.

Bierlaire, M., 2020. A short introduction to PandasBiogeme. (No. TRANSP-OR 200605). Transport and Mobility Laboratory, ENAC, EPFL.

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.

Burgdorf, C., Mönch, A., Beige, S., 2020. Mode choice and spatial distribution in long-distance passenger transport – Does mobile network data deliver similar results to other transportation models? Transp. Res. Interdiscip. Perspect. 8, 100254. https://doi.org/10.1016/j.trip.2020.100254

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 trip generation using mobile phone data: A latent demographics approach. J. Transp. Geogr. 76, 276–286. https://doi.org/10.1016/j.jtrangeo.2017.08.020

Bwambale, A., Choudhury, C.F., Hess, S., 2019c. Modelling departure time choice using mobile phone data. Transp. 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, Md.S., 2020. Getting the best of both worlds: a framework for combining disaggregate travel survey data and aggregate mobile phone

data for trip generation modelling. Transportation. https://doi.org/10.1007/s11116-020-10129-5
Caceres, N., Romero, L.M., Benitez, F.G., 2020. Exploring strengths and weaknesses of mobility inference from mobile phone data vs. travel surveys. Transp. Transp. Sci. 16, 574–601. https://doi.org/10.1080/23249935.2020.1720857
Caceres, N., Romero, L.M., Benitez, F.G., 2013. 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.
Calabrese, F., Di Lorenzo, G., Liu, L., Ratti, C., 2011. Estimating origin-destination flows using mobile phone location data. IEEE Pervasive Comput. 10, 36–44. https://doi.org/10.1109/MPRV.2011.41
Daly, A., Fox, J., Rohr, C., 2002. Advanced modeling to overcome data limitations in the Norwegian transport model, in: Publication of: Association for European Transport. Presented at the European Transport Conference 2002MVA, Limited; Association for European Transport.
Danafar, S., Piorkowski, M., Krysczcuk, K., 2017. Bayesian framework for mobility pattern discovery using mobile network events, in: 25th European Signal Processing Conference, EUSIPCO 2017. pp. 1070–1074. https://doi.org/10.23919/EUSIPCO.2017.8081372
De Heer, W., De Leeuw, E., 2002. Trends in household survey nonresponse: A longitudinal and international comparison. Surv. Nonresponse 41, 41–54.
de Montjoye, Y.-A., Gambs, S., Blondel, V., Canright, G., de Cordes, N., Deletaille, S., Engø-Monsen, K., Garcia-Herranz, M., Kendall, J., Kerry, C., Krings, G., Letouzé, E., Luengo-Oroz, M., Oliver, N., Rocher, L., Rutherford, A., Smoreda, Z., Steele, J., Wetter, E., Pentland, A. "Sandy", Bengtsson, L., 2018. On the privacy-conscientious use of mobile phone data. Sci. Data 5, 1–6. https://doi.org/10.1038/sdata.2018.286
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., 1972. Likelihood. The John Hopkins University Presss, Baltimore, Maryland.
European Environment Agency, 2021. Greenhouse gas emissions from transport in Europe [WWW Document]. URL https://www.eea.europa.eu/ims/greenhouse-gas-emissions-from-transport (accessed 4.11.22).
Gariazzo, C., Pelliccioni, A., Bogliolo, M.P., 2019. Spatiotemporal analysis of urban mobility using aggregate mobile phone derived presence and demographic data: A case study in the city of rome, italy. Data 4, 8.
Ghasri, M., Hossein Rashidi, T., Waller, S.T., 2017. Developing a disaggregate travel demand system of models using data mining techniques. Transp. Res. Part Policy Pract. 105, 138–153. https://doi.org/10.1016/j.tra.2017.08.020
Gundlegård, D., 2018. Transport Analytics Based on Cellular Network Signalling Data. 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., Mobile Traffic Analytics 95, 29–42. https://doi.org/10.1016/j.comcom.2016.04.015
Hess, S., 2014. Latent class structures: taste heterogeneity and beyond, in: Hess, S., Daly, A. (Eds.), Handbook of Choice Modelling. Edward Elgar Publishing.
Holmström, A., 2017. The national travel survey, RVU Sverige (Sweden) 2015-2016. Stockholm.
Holmström, A., Wiklund, M., 2015. The national travel survey, RVU Sverige (Sweden) 2011-2014. Stockholm.
Huang, H., Cheng, Y., Weibel, R., 2019. Transport mode detection based on mobile phone network data: A systematic review. Transp. Res. Part C Emerg. Technol. 101, 297–312. https://doi.org/10.1016/j.trc.2019.02.008
Iqbal, Md.S., Choudhury, C.F., Wang, P., González, M.C., 2014. Development of origin–destination matrices using mobile phone call data. Transp. Res. Part C Emerg. Technol. 40, 63–74. https://doi.org/10.1016/j.trc.2014.01.002

Janzen, M., 2019. Simulating Annual Long-Distance Travel Demand (PhD Thesis). ETH Zurich.
Janzen, M., Vanhoof, M., Smoreda, Z., Axhausen, K.W., 2018. Closer to the total? Long-distance travel of French mobile phone users. Travel Behav. Soc. 11, 31–42.
Kalatian, A., Shafahi, Y., 2016. Travel Mode Detection Exploiting Cellular Network Data. MATEC Web Conf. 81, 03008. https://doi.org/10.1051/matecconf/20168103008
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.
Lovisa Indebetou and Alexander Börefelt, 2018. Resvanor invånare 30-49 år i Umeå tätort - Kartläggning med hjälp av ny datainsamlingsmetod hösten 2017. Umeå.
OpenTripPlanner [WWW Document], 2020. URL http://docs.opentripplanner.org/en/latest/ (accessed 9.2.20).
Phithakkitnukoon, S., Sukhvibul, T., Demissie, M., Smoreda, Z., Natwichai, J., Bento, C., 2017. Inferring social influence in transport mode choice using mobile phone data. EPJ Data Sci. 6, 11. https://doi.org/10.1140/epjds/s13688-017-0108-6
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, Transport Survey Methods in the era of big data:facing the challenges 32, 155–166. https://doi.org/10.1016/j.trpro.2018.10.029
Stopher, P.R., Greaves, S.P., 2007. Household travel surveys: Where are we going? Transp. Res. Part Policy Pract., Bridging Research and Practice: A Synthesis of Best Practices in Travel Demand Modeling 41, 367–381. https://doi.org/10.1016/j.tra.2006.09.005
Tolouei, R., Psarras, S., Prince, R., 2017. Origin-Destination Trip Matrix Development: Conventional Methods versus Mobile Phone Data. Transp. Res. Procedia, Emerging technologies and models for transport and mobility 26, 39–52. https://doi.org/10.1016/j.trpro.2017.07.007
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. Transp. Res. Part C Emerg. Technol., Big Data in Transportation and Traffic Engineering 58, 162–177. https://doi.org/10.1016/j.trc.2015.04.022
Trafikanalys, 2021a. Resvanor.
Trafikanalys, 2021b. Luftfart.
Trafikanalys, 2018. Metodval inför kommande resvaneundersökningar (No. PM 2018:10). Trafikanalys, Stockholm.
Trafikanalys, 2011. Bantrafik.
Trafikanalys, 2018a. Bantrafik.
Trafikanalys, 2018b. Körsträckor.
Trafikverket, 2020. Analysmetod och samhällsekonomiska kalkylvärden för transportsektorn: ASEK 7.0.
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
Wang, H., Calabrese, F., Di Lorenzo, G., Ratti, C., 2010. Transportation mode inference from anonymized and aggregated mobile phone call detail records, in: IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC. pp. 318–323. https://doi.org/10.1109/ITSC.2010.5625188
WSP Analysis and Strategy, 2012. Sampers documentation. Dokumentation av framtagande av kalibreringsmatriser för HHT-projektet. Stockholm.
Yang, M., Pan, Y., Darzi, A., Ghader, S., Xiong, C., Zhang, L., 2022. A data-driven travel mode share estimation framework based on mobile device location data. Transportation 49, 1339–1383. https://doi.org/10.1007/s11116-021-10214-3

9 APPENDIX A - CALCULATION OF INDICATORS IN TABLE 1

The home antenna of a user is computed as follows:

1. Select all stops of the user between 14:00 and 6:00
2. Group these stops by antenna
3. Calculate the total time spent per antenna for these stops
4. The home antenna is the one with highest time spent

A stop is when the user does not move more than two kilometres for more than two hours. If there are no stops during the night, the indicator regular home/night location in Table 1 is set to False, otherwise it is True.

The employment indicator is computed as follows:

1) Select all stops of the user which fulfil all the following criteria:
   a) The stop antenna coverage does not overlap with home antenna
   b) The start time is after 6:00
   c) The end time is before 18:00
   d) The day is Monday-Friday
   e) The stop duration is at least 3h
2) Group the above stops per antenna and calculate
   a) Number of visiting days
   b) Total time spent
3) Work location is the antenna that fulfils all the following criteria:
   a) has at least 2 visiting days (for one week dataset)
   b) no other antenna has a higher total time

10 APPENDIX B – DERIVATION OF LOG LIKELIHOOD ADJUSTMENT

In model specification I and III the log-likelihood (LL) is calculated as a sum of log-likelihoods for modes air, rail, and road, while for model specification II the LL is calculated as a sum of the LL for modes air, rail, car, and bus. The data used in estimating models I and III is extended for model II by adding 'data' obtained by random allocation of road choices to bus and car, so that the LL is calculated over the extended data.

To get 'log-likelihood' values for specifications I and III that are comparable to specification II we must correct models I and III by adding a term corresponding to the prediction of the choice of bus and car, conditional on the choice of road. This Appendix shows how to calculate that term.

The LL we want for the four alternatives is

$$LL = \sum_n \log p_{i,c^*} = \sum_{m=\{air,rail,car,bus\}} \sum_{i,c^*=m} \log p_{i,m}$$

$$= \sum_{i,c^*=air} \log p_{i,air} + \sum_{i,c^*=rail} \log p_{i,rail} + \sum_{i,c^*=car} \log p_{i,car} + \sum_{i,c^*=bus} \log p_{i,bus}$$

where i runs over observations

p_{i,c^*} gives the probability of the observed choice c^* for observation i

$i,c^* = m$ selects the observations that have chosen mode m

In specification II this LL is optimised in the estimation. However, for specifications I and III we do not get $p_{i,\,car}$ and $p_{i,bus}$ directly. Since car is always considered to be available, there are two cases:

Bus not available: $p_{i,car} = p_{i,road}$

$p_{i,bus} = 0$

Bus available $p_{i,car} = p_{i,road}p_{i,car|road} \approx p_{i,road}p_{car|road,a}$

$p_{i,bus} = p_{i,road}p_{i,bus|road} \approx p_{i,road}p_{bus|road,a}$

with $p_{m|road,a}$ calculated as an overall average and therefore not dependent on i. As before, the subscript a indicates that bus is an available alternative and here, *na* denotes bus not available. Then we get

$$\sum_{i,c^* = car} \log p_{i,car} \approx \sum_{i,c^* = car} \log (p_{i,road}p_{car|road})$$

$$= \sum_{i,c^* = car,na} \log p_{i,road} + \sum_{i,c^* = car,a} (\log p_{i,road} + \log p_{car|road,a})$$

$$= \sum_{i,c^* = car} \log p_{i,road} + \sum_{i,c^* = car,a} \log p_{car|road,a}$$

$$= \sum_{i,c^* = car} \log p_{i,road} + N_{car,a} \log p_{car|road,a}$$

where $N_{car,a}$ is the number of people choosing car with bus available. Similarly,

$$\sum_{i,c = bus} \log p_{i,bus} \approx \sum_{i,c^* = bus} \log p_{i,road} + N_{bus} \log p_{bus|road,a}$$

Then we can calculate the adjusted LL for four alternatives as

$$LL = \sum_{i,c^* = air} \log p_{i,air} + \sum_{i,c^* = rail} \log p_{i,rail} + \sum_{i,c^* = car} \log p_{i,car} + \sum_{i,c^* = bus} \log p_{i,bus}$$

$$= \sum_{i,c^* = air} \log p_{i,air} + \sum_{i,c^* = rail} \log p_{i,rail} + \sum_{i,c^* = bus\ or\ car} \log p_{i,road}$$

$$+ N_{car,a} \log p_{car|road,a} + N_{bus} \log p_{bus|road,a}$$

The first three terms here are what comes out of the estimation of specification I and III, while the last two terms, the adjustment, can be estimated by

$$N_{car,a} \log p_{car|road,a} + N_{bus} \log p_{bus|road,a} \approx$$
$$N_{road,a} (p_{car|road,a} \log p_{car|road,a} + p_{bus|road,a} \log p_{bus|road,a})$$

For model III, these conditional choice probabilities could be taken from the model but we judged that the better option is to take them from an external source, namely the NTS of 2011-2016, which means they can also be applied to model I.

11 APPENDIX C - SUMMARY OF MODEL SPECIFICATIONS

Table C1: Summary of model specifications.

Specification:	Description:
I	Joint road utility function for bus and car trips. Variables are weighted according to their relative mode shares from the 2011-2016 NTS. No separation of private and business travellers.
II	Random assignment to either car or bus for all road trips that have an available bus connection. Probability of assignment proportional to the mode shares of the 2011-2016 NTS. No separation of private and business travellers.
III	Road trips assigned to a road nest, which contains the modes bus and car. No separation of private and business travellers.
IV	Business dummy added to the car utility function. The business dummy is true if both the employment indicator and business departure time from Table 1 are true. Based on the road nest structure of specification III.
V	Segmentation into one private model and one business model. Separation based on the same business dummy as in specification IV. Based on the road nest structure of specification III.
VI	Latent class model in which the class probability is formulated as a logit function containing the same business dummy as in specification IV (and a constant for scaling purposes). Based on the road nest structure of specification III.
VII	Latent class model in which the class probability is formulated as a logit function containing the day trip and regular home/night location variables (and a constant for scaling purposes). Based on the road nest structure of specification III.