Long-distance mode choice estimation on joint travel survey and mobile phone network data

VTI Working Paper 2024:1

Angelica Andersson[1,2], Ida Kristoffersson[1], Andrew Daly[3], and Maria Börjesson[4,5]

[1]Traffic Analysis and Logistics, Swedish National Road and Transport Research Institute (VTI)
[2]ITN, Linköping University, Sweden
[3]ITS, University of Leeds, United Kingdom
[4]Transport Economics, Swedish National Road and Transport Research Institute (VTI)
[5]IEI, Linköping University, Sweden

Abstract

The accuracy of a transport demand model's predictions is inherently limited by the quality of the underlying data. This issue has been highlighted by the decline in response rates for transport surveys, which have traditionally served as the primary data source for estimating transport demand models. At the same time, mobile phone network data, not requiring active participation from subjects, have become increasingly available. However, some key trip and traveller characteristics enhancing the prediction power of the estimated models are not collected in mobile phone network data. In this paper we therefore investigate what can be gained from combining mobile phone network data with travel survey data, using the strengths of each data source, to estimate long-distance mode choice models. We propose and estimate a set of mode choice demand models on joint mobile phone network data and travel survey data. We show that combining the two data sources produces more credible estimates than models estimated on each data source separately. The travel survey should preferably include the variables: travel party size, cars per household licence, licence holding, in addition to origin, destination, mode, trip purpose, age, and gender of the respondent.

Keywords

Data combination, Discrete choice modelling, Latent class model, Longdistance mode choice, Mobile phone network data, Transport elasticities, Transport planning, Planning practice

JEL Codes

R41; R42

Long-distance mode choice estimation on joint travel survey and mobile phone network data

Angelica Andersson*[1,2], Ida Kristoffersson[1], Andrew Daly[3], and Maria Börjesson[1,4]

[1]VTI National Road and Transport Research Institute, Sweden
[2]ITN, Linköping University, Sweden
[3]ITS, University of Leeds, United Kingdom
[4]IEI, Linköping University, Sweden

Abstract

The accuracy of a transport demand model's predictions is inherently limited by the quality of the underlying data. This issue has been highlighted by the decline in response rates for transport surveys, which have traditionally served as the primary data source for estimating transport demand models. At the same time, mobile phone network data, not requiring active participation from subjects, have become increasingly available. However, some key trip and traveller characteristics enhancing the prediction power of the estimated models are not collected in mobile phone network data. In this paper we therefore investigate what can be gained from combining mobile phone network data with travel survey data, using the strengths of each data source, to estimate long-distance mode choice models. We propose and estimate a set of mode choice demand models on joint mobile phone network data and travel survey data. We show that combining the two data sources produces more credible estimates than models estimated on each data source separately. The travel survey should preferably include the variables: travel party size, cars per household licence, licence holding, in addition to origin, destination, mode, trip purpose, age, and gender of the respondent.

Keywords: Data combination, Discrete choice modelling, Latent class model, Long-distance mode choice, Mobile phone network data, Transport elasticities, Transport planning, Planning practice.

1 Introduction

In this paper we estimate a large-scale mode choice demand model for long-distance trips in Sweden, on joint mobile phone network data and survey data. To the authors' knowledge, such a joint estimation of a full-scale mode choice model has not previously been published. Mobile phone network data and survey data can each be utilised independently to estimate mode choice models, yet they each come with distinct challenges. By combining the two, we propose a way forward to provide reliable transport forecasting models in the face of falling response rates.

Declining response rates to travel surveys have been a problem for decades, reducing the reliability of transport demand forecasts. According to Stopher & Greaves (2007) there are certain types of households which are less likely to respond to surveys due to time constraints, and this has implications for representativeness. Consequently, valuations of travel time are likely not to be representative when derived from surveys with low response rates. Andersson et al. (2022, Preprint) show that mode shares in recent surveys are biased and a plausible explanation for this is the high non-response rate. Moreover, efforts to control the increasing recruitment costs have occasionally led to respondent recruitment in the pool of individuals who have responded to other surveys before, which in practice exacerbates the problem with unrepresentative survey respondents (Wang et al., 2022). GPS tracking surveys, proposed to reduce reporting errors, unfortunately often suffer from even lower response rates (McCool et al., 2021), sometimes even lower than 10% (Indebetou

& Börefelt, 2018), possibly for integrity or technical reasons. However, small GPS datasets have proven useful for validation of model development for other data sources, such as in Bwambale et al. (2019), where GPS data are used as comparison for the model results of departure time choice produced using mobile phone network data. For these reasons, i.e. decreasing response rates, weaker representativeness, and increasing costs of surveys, efforts have been made to find complements to survey data for transport demand model estimation.

Bwambale et al. (2019) proved that mobile phone network data on its own can be useful for econometric trip generation modelling, while Andersson et al. (2022, Preprint) succeeded in estimating a mode choice logit model on mobile phone network data only, finding that the model produced reasonably credible parameter estimates. There are, however, issues reducing the accuracy of the transport model estimated only on mobile phone network data, for instance that the trip purpose is not stated and that it can be difficult to separate car, bus and truck trips, which use the same road infrastructure. Bwambale et al. (2019) inferred the most probable trip purpose from mobile phone usage patterns inserted into a latent class structure. Still, there were remaining uncertainties in the trip purpose. And, in many cases, ours included, the mobile phone network data are too aggregate to apply this method to infer trip purpose. To account for trip purpose in the mode choice model, Andersson et al. (2022, Preprint) applied a latent class model using location and time as indicators to identify classes of trips interpreted as private and business trips. They also applied a nested logit structure to estimate separate parameters for bus and car trips in the mode choice model.

The objective of this paper is to combine mobile phone network data with travel survey data, leveraging the strengths of each data source to enhance the accuracy of the resulting mode choice model. The use of travel survey data allows for the addition of variables that are only available in the survey data. In the travel survey data, the trip purpose and road travel mode are known, but also variables such as party size, driving licence and car access, which have the potential to increase the forecasting model precision. On the other hand, collecting large amounts of mobile phone network data is relatively cheap, and another advantage is that it is gathered without requiring active participation from travellers, eliminating the issue of low response rates.

We investigate i) how to combine mobile phone and travel survey data in large scale mode choice model estimation, ii) whether the joint model produces more credible estimates than when models are estimated on the two data sources separately. Furthermore, we also provide recommendations for data collection and combination for future mode choice demand models and analyse policy implications of the estimated results.

To our knowledge, no previous study has utilised mode choice information from mobile phone network data and survey data in a joint mode choice model estimation. However, passive data, not requiring active participation from travellers, have been combined with active survey data in the past. Janzen (2019) combined mobile phone network data with survey data using Histogram Matching at the origin-destination level, picking mode choices directly from the survey data to create a synthetic population. Brederode et al. (2019) used a multi-proportional gravity model to fuse the two data sources to obtain total travel demand, also picking mode choices from survey data. Bwambale et al. (2020) combined mobile phone network data and survey data for trip generation purposes, finding an improved spatial and temporal transferability in the combined model compared to the base model based only on survey data. Zhang et al. (2019) also combined active and passive data in an approach which is similar to ours, but in their case the passive data consisted of a combination of transit card and a car GPS tracker data rather than mobile phone network data.

2 Data

Three data sources are used in this paper. First, travel supply data (such as travel times, travel costs, number of transfers etc.) were provided by The Swedish Transport Administration. Second, we use the Swedish National Travel Surveys (NTS) from 2011-2016. Third, we use mobile phone network data from one week of travel in 2018 which have been pre-processed to include origin zone, destination zone, likely travel mode, and some other time and location-based indicators. Descriptive statistics of the datasets can be found in Appendix A.

2.1 Supply data

We use the traffic analysis zones (TAZs) that the Swedish Transport Administration applies for long-distance trips. There are 682 zones in total, which means that each zone covers on average 770 km^2 and 15,000 inhabitants, although the sizes of the zones vary depending on land use density (see Appendix B for a map of the zones). We consider only domestic trips longer than 100 km. All supply data (travel times, travel costs, number of transfers etc.) use 2017 as reference year.

2.2 Travel survey data

The National Travel Surveys (NTS) from 2011-2016 had an average response rate of 41% (Holmström, 2017; Holmström & Wiklund, 2015). Together they include 20,069 trip observations, after discarding incomplete observations. The annual samples are independent, i.e., they are not panel surveys. Respondents in the NTS data are randomly selected from a stratified sample of the Swedish population, based on the respondent's county, age group, and gender. The respondents were contacted by telephone and asked to report trips made during one pre-determined measurement day. The measurement days were randomly spread throughout the year. Additionally, respondents were asked to report trips longer than 100 km during the previous month and trips exceeding 300 km in the past three months. Trips exceeding 100 km constitute approximately 20 percent of the total vehicle kilometres travelled in Sweden.

The NTS data contain information about the trip and traveller, such as trip purpose, gender, age, travel party size, licence holder, cars in household and so on. However, they are probably not representative of the trips made by the full population. Analysis of the corresponding 2005-2006 NTS data indicates that the distribution of the distances of car trips differs between the measurement day and the measurement month (WSP, 2012). This is presumably because "short" long-distance trips by car are often forgotten after some time. On the other hand, the long-distance trips reported for the measurement day are in total significantly fewer than what is indicated by the monthly travel (ibid), possibly because those making a long-distance trip on the measurement day were more difficult to reach. Since the 2005-2006 survey the response rates have fallen sharply (which has been a problem for many travel surveys worldwide). There are also indications of declining representativeness in the more recent travel surveys. Andersson et al. (2022, Preprint) noticed that the more recent NTS data used in the present study (2011-2016) have unreasonably high shares of air trips. In the 2018 mobile phone network data, 2% of the trips are made by air, while this share is 15% in the 2011-2016 NTS data and 4% in the 2005-2006 NTS data[1].

[1] When adjusting all data sources to 2018 levels using annual ticket sales reported by air and train operators, and total car mileage.

2.3 Mobile phone network data

The mobile phone network data consist of 92,011 trip observations which have been derived from raw data by applying the "Route/Antenna method" described in Breyer et al. (2021). This method produces an anonymised table of trips containing origin zone, destination zone, and likely travel mode. It also includes a few time- and location-based indicators, such as whether the outbound and return trip is conducted on the same day (daytrip) and whether the traveller has a regular home/night location. The raw data is so called xDR[2] data, which provides a location and timestamp for the connected mobile phone network antenna each time a call is made, a text is sent, the internet is used (by the traveller or by apps or operating system working in the background) or whenever a handover event to a new antenna occurs. Analysis by Breyer et al. (2021) showed that antenna position updates normally occur within 5 minutes. The spatial resolution is limited by the coverage range of the antenna. Antenna ranges vary due to several different factors, but as a rule of thumb the coverage range is in the order of 10 km in rural areas and 500 m in urban areas. Because the TAZs are larger than the coverage area of the antennas and exhibit similar variations with population density, this data is well-suited for estimating a long-distance transport model. A higher level of resolution in urban areas can be expected eventually as high-band 5G antennas are introduced in larger numbers, possibly increasing the scope of the methods used in this paper to also include regional transport models. The Route/Antenna method used to derive trips is based on the proximity between antennas and built transport infrastructure, which means that bus and car trips are both labelled as road trips in the mobile phone network data.

If the main issue with survey data is its representativeness of the entire population, a key concern in this paper is whether mobile phone network data are more representative of the travel patterns of the full population compared to the travel survey. There are reasons to believe that this is the case. First, the mobile phone network operator claims that they do not target any specific socio-economic group in their sales. Second, mobile phone network data do not experience a drop in response rates from travellers with high value of time (who do not respond to surveys due to time constraints). However, a counterargument against the high representativeness of mobile phone network data is that some socio-economic groups may be more likely to have dual SIM-cards than others, which increases their chances of being sampled.

3 Methodology

3.1 Model specifications

In this paper, four mode choice models are estimated and compared (Model I – Model IV) using Python Biogeme (Bierlaire, 2020). Model I is the preferred model specification estimated on mobile phone network data only, which is defined and estimated in Andersson et al. (2022, Preprint)[3]. Model II applies the same model specification as Model I but is estimated on joint survey and mobile phone network data. Model III improves the model specification of Model II. Model IV adds four variables which are included in the survey data only.

Model I is a nested logit mode choice model for long-distance passenger travel, including the four modes air, bus, car, and train. The utility functions for private (p) trips are

[2]xDR data is an extension of the previous Call Detail Record (CDR) data which are collected for billing purposes by mobile phone network operators. While CDR data are recorded for each call or text message, xDR data are also recorded for active or passive internet usage.

[3]The peak-hour dummy variable had to be excluded as it is missing in the survey data and the utility functions of Model I and II must be identical in order to enable a fair comparison.

$$V_{i,p} = \kappa_{i,p} + \chi_{i,p} t \mathbb{1}_{\{\mathsf{b},\mathsf{a},\mathsf{r}\}}(i) + \beta_p c + \alpha_{i,p} \sigma \mathbb{1}_{\{\mathsf{r}\}}(i) + \omega_p \log(\mathsf{w}) \mathbb{1}_{\{\mathsf{f},\mathsf{b},\mathsf{r}\}}(i) + \zeta_p z \mathbb{1}_{\{\mathsf{a}\}}(i), \quad (1)$$

and the business (b) utility functions are

$$V_{i,b} = \kappa_{i,b} + \chi_{i,b} t \mathbb{1}_{\{\mathsf{f},\mathsf{a},\mathsf{r}\}}(i) + \beta_b c + \alpha_{i,b} \sigma \mathbb{1}_{\{\mathsf{f},\mathsf{r}\}}(i) + \omega_b \log(\mathsf{w}) \mathbb{1}_{\{\mathsf{f},\mathsf{r}\}}(i) + \eta_{i,b} n \mathbb{1}_{\{\mathsf{f},\mathsf{r}\}}(i) \quad (2)$$

where $i \in \{\mathsf{f}, \mathsf{b}, \mathsf{a}, \mathsf{r}\}$ denotes the modes air (f), bus (b), car (a) or train (r), and $\mathbb{1}_{\{\mathcal{S}\}}(x)$ is the indicator function, which is 1 if $x \in \mathcal{S}$ and 0 otherwise. Furthermore, κ denotes an alternative specific constant, χ denotes the travel time parameter, t denotes the travel time, β is the cost parameter, c is the travel cost, σ is the sum of the distance between the centroid of the origin zone and the closest mode terminal (airport or train station), plus the distance between the centroid of the destination zone and closest mode terminal, with α as the parameter of this connecting trip representation σ. ω is the parameter belonging to first wait time w (defined as half headway), ζ is the parameter for the dummy variable z, which is 1 if the trip started during a weekend and 0 otherwise and η denotes the parameter for the number of boardings n. The alternative specific constant for car is fixed to zero for both trip purposes, and the bus alternative is not considered to be available for business trips since business long-distance bus trips are rare.

As described in Section 2.3, the mobile phone network data used in our model estimation are derived from xDR raw data. The "Route/Antenna method" developed by Breyer et al. (2021) is applied to identify the trips and their modes. In this identification process, bus and car cannot be distinguished because they use the same infrastructure. Because of this, we place the car and bus alternatives in a road nest. This implementation of the bus and car alternative in a joint nest is inspired by the approach used in Daly et al. (2002). The probability of choosing mode $i \in \{\mathsf{b}, \mathsf{a}\}$ within the road nest is given by

$$P_{road,i} = \Pr(road)\Pr(i|road) = \frac{\exp V^*_{road}}{\sum_k \exp V^*_k} \frac{\exp V_i}{\sum_j \exp V_j} \quad (3)$$

where $V^*_{road} = \theta \log \sum_j \exp(V_j/\theta)$, and V_j is the utility for mode $j \in \{\mathsf{b}, \mathsf{a}\}$ given by 1 (the bus utility is set to $-\infty$ when there is no available bus connection), and $k \in \{\mathsf{road}, \mathsf{f}, \mathsf{r}\}$. In Andersson et al. (2022, Preprint) the nest parameter is estimated to 1 (under the restriction $\theta \in]0, 1]$), so in this paper, the nest parameter is constrained to 1.

The trip purpose is also unknown in the mobile phone network data. The distinction between private and business travellers is important since they have different travel preferences. To take these differences into account, Andersson et al. (2022, Preprint) formulated a latent class model (Hess, 2014), where the probability that a trip belongs to the latent class for business trips is

$$q_b = \frac{1}{1 + e^{-(\gamma + d\gamma_d + h\gamma_h)}}, \quad (4)$$

and the probability that it belongs to the latent class for private trips is

$$q_p = 1 - q_b, \quad (5)$$

where γ, γ_d and γ_h are parameters to be estimated (relating to a constant, a daytrip and a regular home/night location variable respectively; see Appendix C for definition).

In the joint estimation in Model II, we use the same model specification as in Model I. When combining NTS data and mobile phone network data, the mobile phone network data use the latent class structure, while the NTS data use the trip purpose reported by respondents. Similarly, the road nest structure is used for mobile phone network data observations just as in Model I, while the NTS observations are assigned directly to either the bus or car alternative according to the reported mode. To account for differences in the distribution of the error term between the two data sources, we introduce a scaling parameter ϕ, which multiplies all the survey utility functions and is estimated along with the other parameters (see Appendix D for explicit utility functions for model II). This approach is analogous with what has been done to combine revealed preference and stated preference data, as originally proposed by Ben-Akiva & Morikawa (1990), followed by the adaptation to logit models by Daly & Bradley (1991). We also estimate separate alternative-specific constants for the two data sources since the unobserved effects may differ between the two sources, including the possibility that the probability of being observed by different modes may differ in the two data sources.

The specification of Model II is improved in several ways in Model III (see Appendix E for explicit utility functions for model III). First, the travel survey data includes information regarding the traveller's age and the trip party size, which means that the cost estimate can be improved. A larger party size in the car implies that the travel cost can be split among more travellers, and age can be used to adjust prices for young bus travellers enjoying discounts[4]. We use a shared cost parameter for private trips from survey and mobile phone network data, and separate cost parameters for business trips from the two data sources, since the cost parameters were significantly different only for the business trips.

Since the travel purpose is known in the survey data, the accuracy of the parameters not included in the latent class model is improved when these data are added in the model estimation. An improved estimation of these parameters means in turn that the latent class variables estimated on the mobile phone network data only can be estimated to better match the two classes private and business, than when only mobile phone network data is applied. The second improvement in Model III is therefore to replace the old latent class variables, which were found to be insignificant in Model II, with new variables that can identify the trip purpose more effectively when both data sources are used. The first new variable is the trip distance δ. The second new variable is a dummy ρ, being one if the destination zone belongs to the 10% densest zones in the country and zero otherwise. The density indicator of the destination zones is the sum of inhabitants and employments per square km. The new expression for the probability of belonging to the business class thus becomes: $q_b = \frac{1}{1+e^{-(\gamma+\rho\gamma_\rho+\delta\gamma_\delta)}}$, where γ, γ_ρ and γ_δ are parameters to be estimated. Estimation of the same specification as Model III but without survey data resulted in insignificant latent class variables, meaning that the inclusion of survey data is key for finding a good separation of trip purposes.

Model IV is based on Model III, but we add four variables which are available only in the survey data: group g (a dummy which is true if the party size is greater than 8, intended to capture hired buses) on the bus utility function, a female solo dummy F (true if the traveller is female and travelling alone) on the bus and train utility functions, and cars per licence in the household y (restricted to the interval $[0, 1]$ and set to 0 when there are no licences in the household) on the car utility function, with the complementary dummy no licence o indicating whether the traveller has a driving licence or not on the car utility function. See Appendix F for the explicit utility functions of model IV.

The survey data includes panel data in the sense that all trips by the same respondent made during the past month(s) can be linked, while repeated trips made by the same individual cannot be linked in the mobile phone network data. To mitigate the effect of this potential misspecification from the correlation between trips made by the same individual, we report robust t-values in the results section based on the sandwich estimator, rather than the usual t-values.

3.2 Values of time and elasticities

The value of travel time (VTT) is defined as the marginal rate of substitution between time and money

[4] We only have access to youth prices for bus trips, not for other modes.

$$VTT = \frac{\partial u(c,t)}{\partial t} \bigg/ \frac{\partial u(c,t)}{\partial c} \tag{6}$$

for some utility function $u(c,t)$ which depends on travel time t and cost c. For utility functions that are linear in cost and time, this simplifies to $VTT = \chi/\beta$, where χ and β denote the time and cost parameters as defined previously.
The elasticity for individual n is calculated as

$$E_{x_{ink}}^{P_n(i)} = \frac{\partial P_n(i|x_n, C_n)}{\partial x_{ink}} \frac{x_{ink}}{P_n(i|x_n, C_n)}, \tag{7}$$

where x_{jnk} is the k:th variable of alternative i and individual n, $P_n(i)$ denotes the probability that individual n chooses alternative i, and C_n is the choice set of individual n. The aggregate elasticity is then calculated as a weighted average of the elasticities of all individuals using the choice probabilities as weights (Ben-Akiva & Lerman, 1985)

$$E_{x_{ink}}^{\bar{P}_n(i)} = \frac{\sum_n P_n(i) E_{x_{ink}}^{P_n(i)}}{\sum_n P_n(i)} \tag{8}$$

where $\bar{P}_n(i)$ denotes the expected share which chooses alternative i.

4 Results and discussion

In this section we first investigate whether a joint model estimation can help resolve some of the challenges related to an equivalent model estimated solely on mobile phone network data. We then add variables unique to the survey data and investigate whether this addition is valid and to what extent it improves the model. Parameter estimates, resulting values of time and elasticities are all discussed below.

4.1 Estimation results

Table 1 compares Model I, estimated on mobile phone network data only and Model II, adding survey data to the estimation. First, there is an overall improvement in the t-values of most parameters, which is expected since the number of observations increases. There is a large increase in the t-value for the private car time parameter. This is probably because the identification of car versus bus trips was weak in Model I, as there is no definitive information on which of the road trips were car trips and which were bus trips. Second, the scale parameter for the survey data, ϕ, is estimated to 0.837, indicating that the response scale is lower in the survey data. Third, the latent class variables become insignificant in Model II, confirming the concern that the class split identified in Model I does not accurately identify trip purposes between business and private trips.
Since the travel purpose is known in the survey data, the non-latent class parameters are estimated to better fit the two classes private and business when survey data is applied in Model II, than when only mobile phone data is applied in Model I. This can be seen by the finding that the values of time are more consistent with expectations in Model II than in Model I (see Table 3). As survey data is added in Model II, the magnitude of the travel cost parameter increases by a factor of 7, the magnitude of the car time parameter roughly increases by a factor of 2, and the air time parameter increases by a factor of 3 for business trips in Model II. Only the train time parameter stays roughly the same.

Table 1: Comparison between Model I, estimated on mobile phone network data only and Model II estimated on both mobile phone network data and survey data. Subscript m indicates mobile phone network data and s survey data. A full list of abbreviations is provided at the end of the paper.

	Model I				Model II			
	Private		Business		Private		Business	
	Value	Robust t-value	Value	Robust t-value	Value	Robust t-value	Value	Robust t-value
$\kappa_{\mathsf{b},s}$					-2.47	-15.1		
$\kappa_{\mathsf{b},m}$	5.61	9.9			-6.34	-13.4		
$\kappa_{\mathsf{f},s}$					-4.58	-29.8	1.93	2.2
$\kappa_{\mathsf{f},m}$	-5.93	-9.6	-0.412	-1.1	-4.83	-48.5	1.47	4.4
$\kappa_{\mathsf{r},s}$					-1.52	-17.1	-0.944	-2.2
$\kappa_{\mathsf{r},m}$	0.902	6.6	2.672	12.0	-0.196	-4.5	-0.786	-2.4
β	-0.00234	-7.2	-0.000166	-2.3	-0.00103	-24.5	-0.00112	-7.1
α_{f}			-0.0286	-12.4			-0.0522	-12.4
α_{r}	-0.0224	-18.6	-0.0388	-13.7	-0.0223	-44.5	-0.0131	-4.5
ω	-0.135	-4.4	-0.483	-13.1	-0.132	-12.1	-0.199	-2.4
η			-1.09	-11.0			-0.481	-4.5
χ_{b}	-0.0461	-10.4			-0.00761	-20.1		
χ_{a}	-0.0140	-9.7	-0.0144	-29.7	-0.0104	-73.1	-0.0246	-12.8
χ_{f}			-0.0125	-6.0			-0.0410	-8.0
χ_{r}	-0.00236	-3.1	-0.0184	-20.4	-0.00528	-31.9	-0.0143	-9.2
ζ	0.882	12.9			0.535	27.0		
ϕ					0.837	28.9		
γ	-0.0229	-0.2			-2.38	-32.6		
γ_d	0.0803	2.3			-0.0114	-0.2		
γ_h	0.127	4.4			-0.111	-1.8		
BIC	119891.3				137132.9			
LL(β)	-59814				-68399			
Parameters	23				29			
Observations	92011				112080			

Table 2 shows the result of adding variables unique to the survey data (Model IV) compared with an improved variant of the combined model (Model III). Note first that the new latent class variables present in both Model III and Model IV are significant even in the presence of survey trip purpose information, indicating that the trip purpose is more adequately identified by these new latent class variables. This means that by combining the two data sources, we can both identify latent class variables that can more accurately determine trip purposes in mobile phone network data and obtain better estimates of these latent class variables.

Moving on to the addition of variables unique to the survey data (group g on the bus utility function, female solo dummy F on the bus and train utility functions, cars per licence in the household ψ on the car utility function and no licence dummy o), the BIC criterion, which takes into account model fit while penalising by the number of included variables, implies that Model IV, which includes variables unique to the survey data, is preferred over Model III. As notable changes between the two estimation results occur only in the alternative specific constants, we can conclude that our method of data combination is sound, as there is seemingly no correlation between the variables which are only available for survey data and the other variables. As in the Swedish long-distance model estimated on older data, a party size greater than eight increases the probability of choosing bus, females travelling alone are more likely to travel by train or bus, having more cars per licence in the household increases the probability of travelling by car, and not having a licence decreases the probability of travelling by car.

Table 2: Comparison of estimation results for the model with variables unique to the survey data (Model IV) with an improved variant of the combined model (Model III). Subscript m indicates mobile phone network data and s survey data. A full list of abbreviations is provided at the end of the paper.

	Model III				Model IV			
	Private		Business		Private		Business	
	Value	Robust t-value	Value	Robust t-value	Value	Robust t-value	Value	Robust t-value
$\kappa_{\mathsf{b},s}$	-2.30	-14.8			-1.69	-9.8		
$\kappa_{\mathsf{b},m}$	-5.72	-9.7			-6.16	-9.6		
$\kappa_{\mathsf{f},s}$	-4.19	-29.2	0.658	0.7	-2.95	-22.3	1.42	1.6
$\kappa_{\mathsf{f},m}$	-5.56	-27.8	0.451	1.3	-5.58	-28.2	0.709	2.1
$\kappa_{\mathsf{r},s}$	-1.48	-17.0	-0.921	-2.3	-0.618	-6.8	-0.404	-0.9
$\kappa_{\mathsf{r},m}$	-0.0116	-0.3	-1.317	-3.9	-0.021	-0.5	-1.10	-3.3
ψ					2.04	17.3	1.25	4.2
β	-0.00127	-24.7			-0.00124	-23.6		
β_s			0.000101	0.5			0.000194	1.0
β_m			-0.00151	-9.4			-0.00153	-9.3
α_{f}			-0.0442	-12.0			-0.0445	-12.1
α_{r}	-0.0215	-48.4	-0.0109	-4.0	-0.0211	-48.1	-0.0109	-3.9
ι					0.915	13.1		
ω	-0.126	-11.9	-0.197	-2.7	-0.127	-11.9	-0.191	-2.6
η			-0.465	-4.3			-0.414	-3.8
λ					-1.17	-6.9	-0.801	-1.3
μ					5.96	18.9		
χ_{b}	-0.00753	-20.2			-0.00777	-18.4		
χ_{a}	-0.00954	-53.9	-0.0225	-10.7	-0.00961	-54.7	-0.0235	-11.0
χ_{f}			-0.0306	-7.0			-0.0315	-7.1
χ_{r}	-0.00457	-26.5	-0.0135	-6.6	-0.00465	-26.8	-0.0140	-6.9
ζ	0.492	27.0			0.494	27.1		
ϕ	0.905	27.4			0.901	26.5		
γ	-3.66	-39.2			-3.67	-39.3		
γ_ρ	-0.449	-3.1			-0.444	-3.1		
γ_δ	0.00278	19.9			0.00277	19.8		
BIC	136531.4				134850.1			
LL(β)	-68092				-67217			
Parameters	30				36			
Observations	112080				112080			

4.2 Comparison of values of time

Values of time are presented in Table 3 for Model I and Model II, and in Table 4 for Model III and Model IV.

Table 3: Values of time for Model I and Model II in SEK/h (10SEK ≈ 1€ at reference time 2017).

Mode	Private I	Business I	Private II	Business II
Car	359	5178	603	1317
Bus	1180	-	441	-
Train	60	6628	306	765
Air	-	1180	-	2198

Values of time for Model III and Model IV are similar since the main differences between the models are in the alternative specific constants and in the variables unique to survey data. The comparison of the results in Tables 3 and 4 demonstrates that large gains were made by combining the two data sets, in terms of behavioural consistency and credibility of the estimates. In Model I, which is estimated on mobile phone network data only, the valuation of private bus travel time is higher than expected. When the data is combined in Model II the values of time are more credible, but both private bus and car values of time are still higher than expected. This issue is reduced when more accurate price information available in the survey data is used in Model III (based on party size and traveller age). In Model III and Model IV, all business values of time are higher than their respective private values of time for the same mode, as expected.

Table 4: Values of time for Model III and Model IV in SEK/h (10SEK ≈ 1€ at reference time 2017).
*Business values of time are only based on the mobile phone network data cost parameter as the survey data cost parameter became insignificant.

Mode	Private III	Business* III	Private IV	Business* IV
Car	452	892	465	918
Bus	357	-	376	-
Train	217	534	225	547
Air	-	1213	-	1234

The values of time in Model IV are roughly three times higher than the values in the Swedish CBA guidelines (Swedish Transport Administration, 2020). The latter are based on the values estimated by Börjesson & Eliasson (2014) on stated preference (SP) data[5], while our models are estimated on revealed preference (RP) data. Revealed preference data are known to result in higher values of time than stated preference data (Börjesson, 2008; Brownstone & Small, 2005), for two possible reasons. First, since stated preference is a reported behaviour, there is a risk that respondents fail to fully account for time constraints during the survey trip, which implies that the values of time are underestimated. Second, there is typically more error in the cost variable than in the time variable (Varela et al., 2018) in RP data, leading to attenuation bias in the cost parameter and consequently an overestimation of the values of time. In this paper, we observed that the attenuation bias of the cost parameter decreased as the price variable was adjusted for party size and youth ticket prices in Model III. Therefore, one of the most significant improvements that can be

[5]The Swedish Transport Administration (2020) use 2017 as reference year and their values are based on the findings of Börjesson & Eliasson (2014) who use stated choice data from 2007-2008.

made to RP data (whether from surveys or mobile phone network data) is to enhance the accuracy of the cost variable.
Furthermore, in Model IV, the value of train travel time is lower than the value of bus travel time, which contrasts with Börjesson & Eliasson (2014). This difference is not surprising, given that in SP data, the bus and train values were estimated based on data from respective user groups (bus users and train users), while in our present model, the values of time for both bus and train are estimated for all travellers. Since we anticipate that travellers with higher values of time self-select into the train (which is faster and more expensive), and those with lower values of time self-select into the bus (which is slower and cheaper), we expect the train values of time to be higher in the SP data, but not in the RP data, all else being equal.

4.3 Comparison of elasticities

Time and cost own-elasticities for the final model (IV) are presented in Table 5. Own-elasticities for private air travel time, business air and car travel time, and private train and car travel cost are similar to the elasticities used in the current Swedish long-distance model (Beser & Algers, 2002) and other comparable studies (Börjesson, 2014; Kristoffersson & Liu, 2023). The most notable differences occur in the own-elasticities for private bus travel time, private air travel cost and business train travel time. The private bus travel time elasticity is higher in our model (-1.79) than in Beser & Algers (2002) (-0.66). The private air travel cost elasticity is also higher in our model (-1.28) than in Beser & Algers (2002) (-0.40). The business train travel time elasticity, on the other hand, is notably smaller in our model (-0.23) than in Beser & Algers (2002) (-1.32), Börjesson (2014) (-2.1), and Kristoffersson & Liu (2023) (-1.1). Furthermore, the private car travel time elasticity is more elastic in our model (-0.62) than in Beser & Algers (2002) (-0.28), while the opposite is true for the private bus travel cost elasticity (-0.18 compared to -0.42).

Table 5: Trip own-elasticities with respect to travel time and travel cost for the final model (Model IV). Only private own-elasticities with respect to cost are reported as the business cost parameter from the survey data became insignificant.

	Travel time		**Travel cost**
Mode	**Private**	**Business**	**Private**
Train	-0.55	-0.23	-0.60
Air	-	-0.80	-1.28
Car	-0.62	-0.96	-0.12
Bus	-1.79	-	-0.18

Cross-elasticities are presented in Table 6. The cross-elasticities are generally small, but not as small as previous literature suggests (Beser & Algers, 2002; Rich & Mabit, 2012), with the exception of the cross-elasticity of train in response to changes in bus travel time, which is less elastic in our model. The business air travel modal shift due to changes in train travel time even turns elastic. Private train and air trips with respect to car travel time also turn elastic in our model. The cross-elasticities for business train and car trips with respect to air time, business train trips with respect to car travel time, private air and car trips with respect to bus travel time are all very similar to the elasticity values reported by Beser & Algers (2002) and/or Rich & Mabit (2012).
There is a wide variation in reported elasticities in previous literature. This is expected since elasticities vary between contexts, for instance with modal shares. Rich & Mabit (2012) explicitly provide different elasticities for different trip lengths, Kristoffersson & Liu (2023) consider international trips, meaning on average longer trip lengths than in our case, and Beser & Algers (2002) and Börjesson (2014) consider the same trip lengths as in this study, but based on older survey data. The elasticities produced by a model also

depend on the choice dimensions it incorporates, for instance, if it is a mode choice model only or a combined mode and destination choice model. The trip elasticity for a given mode produced by the mode choice model is expected to be higher than the trip elasticity for the same mode produced by the mode-destination choice model. This is simply because in the latter model, travellers can adapt by changing their destination instead of changing their mode. If frequency is also modelled, travellers can adapt by staying at home, implying that trip elasticities for a given mode increase. Among the studies we compare our results to, Rich & Mabit (2012) and Kristoffersson & Liu (2023) are mode-destination choice models while Beser & Algers (2002) and Börjesson (2014) model trip mode, destination and frequency.

Table 6: Trip cross-elasticities with respect to travel time for the final model (Model IV).

Mode	**Changed variable**	**Private**	**Business**
Air	Train time	0.44	1.07
Car	Train time	0.26	0.43
Bus	Train time	0.24	-
Train	Air time	-	0.11
Car	Air time	-	0.09
Train	Car time	1.06	0.11
Air	Car time	1.48	0.19
Bus	Car time	0.84	-
Air	Bus time	0.06	-
Train	Bus time	0.01	-
Car	Bus time	0.15	-

5 Conclusions and policy implications

In this paper we have investigated to what extent adding survey data to mobile phone network data for mode choice demand model estimation can mitigate the challenges of mobile phone network data, as well as to what extent the estimation can be improved by adding variables unique to the survey data.

Both survey data and mobile phone network data are associated with benefits and limitations. Mobile phone network data is limited in its lack of ground truth for the bus and car split, in the lack of socio-economic information about the traveller, as well as in the lack of information about trip purpose. More recent survey data, on the other hand, often has too few observations to obtain significance in important parameters, and it is unlikely that they capture a representative sample of the population. In this paper we find that by combining the two datasets in a mode choice model, we get more credible model parameters than from separate estimation.

As we have shown that mobile phone network data and survey data complement each other, it is advisable to continue to collect travel survey data. To complement mobile phone network data in mode choice model estimation, we have shown that the survey data should preferably include trip purpose, travel party size, and the socio-economic variables cars per licence in household, driving licence, age, and gender of the respondent. The mobile phone network data used for model estimation need to contain origin zone, destination zone and mode, or enough detail to extract such information (for example raw xDR data). Furthermore, the possibility of accurately predicting travel mode is higher if the data also contain information on whether the trip was conducted on a weekend or weekday. The number of trip observations is of great importance, and in case the survey fails to capture enough responses to obtain significance in important parameters unique to the survey data, one option is to add survey data from several consecutive years to increase the sample size. On the supply data side, it is necessary to have accurate travel costs, travel times, headway, locations of airports and train stations and the number of transfers per mode. Specific efforts should be placed on the collection of accurate cost data, as inaccurate cost data has been shown to cause an over-estimation of values of travel time (Kristoffersson et al., 2022).

The method presented in this paper can be used to combine the two data sources. Furthermore, any additional variables from the survey data may be included in the model as long as the effect of these additional variables is absorbed by the alternative-specific constants in the survey and mobile phone network data utility functions. A prerequisite for this is that the additional variables are not correlated with the existing variables.

Most of the elasticity results in this paper are on par with previous studies; however, our results suggest higher business air cross-elasticities in response to business train travel time than has previously been observed, as well as higher private train and air cross-elasticities in response to car travel time.

At present, mobile phone network data is applicable mainly for domestic long-distance mode choice models, which suffer from few survey observations and do not have high demands of high spatial detail. However, if response rates continue to decline and high-band 5G antennas become more widespread in urban areas, this method of data combination may also become appropriate for regional or urban models.

In this paper, we have shown that the proposed method of data combination is indeed valid, and that by combining the two data sources we can mitigate the limitations of each separate data source and thus maintain good quality mode choice demand forecasting models, even in the face of declining survey response rates.

Acknowledgements

This paper has been funded by the Swedish Transport Administration within the scope of the research project Demopan-2 under Grant TRV 2021/22422.

List of abbreviations

a	car
b	bus
c	cost
d	daytrip variable
E	Elasticity
f	air
F	female solo variable
g	group
h	home
m	mobile phone network data
n	number of transfers
o	no license variable
p	private
P	probability
q	probability (latent class)
r	train
s	survey data
t	travel time
u	utility
V	systematic utility
w	first wait
x	business
y	cars per licence variable
z	weekend variable
α	connecting trip parameter
β	cost parameter
γ	latent class parameters
δ	trip distance variable
ζ	weekend parameter
η	number of transfers parameter
θ	nest parameter
ι	female solo parameter
κ	alternative specific constant
λ	no licence parameter
μ	group parameter
ρ	density variable
σ	connecting trip variable
ϕ	scaling parameter
χ	time parameter
ψ	cars per licence parameter
ω	first wait parameter

References

Andersson, A., Börjesson, M., Breyer, N., Daly, A., Engelson, L., & Kristoffersson, I. (2022, Preprint). *Mode Choice Latent Class Estimation on Mobile Network Data. SSRN.* Retrieved 2023-02-28, from https://doi.org/10.2139/ssrn.4246865

Ben-Akiva, M., & Lerman, S. R. (1985). *Discrete choice analysis: theory and application to travel demand* (No. 9). Cambridge, Massachusetts London: The MIT Press.

Ben-Akiva, M., & Morikawa, T. (1990). Estimation of switching models from revealed preferences and stated intentions. *Transportation Research Part A: General*, *24*(6), 485–495. Retrieved from https://doi.org/10.1016/0191-2607(90)90037-7

Beser, M., & Algers, S. (2002). SAMPERS — The New Swedish National Travel Demand Forecasting Tool. In L. Lundqvist & L.-G. Mattsson (Eds.), *National Transport Models: Recent Developments and Prospects* (pp. 101–118). Berlin, Heidelberg: Springer. Retrieved 2023-09-12, from https://doi.org/10.1007/978-3-662-04853-5_9 doi: 10.1007/978-3-662-04853-5_9

Bierlaire, M. (2020). *A short introduction to PandasBiogeme.* (Tech. Rep. No. TRANSP-OR 200605). Transport and Mobility Laboratory, ENAC, EPFL. Retrieved from http://transp-or.epfl.ch/documents/technicalReports/Bier20.pdf

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)* (pp. 1–8). Retrieved from 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. *Transportation Research Procedia*, *52*, 211–218. Retrieved from https://doi.org/10.1016/j.trpro.2021.01.024 (Publisher: Elsevier)

Brownstone, D., & Small, K. A. (2005). Valuing time and reliability: assessing the evidence from road pricing demonstrations. *Transportation Research Part A: Policy and Practice*, *39*(4), 279–293. Retrieved 2022-04-20, from https://www.sciencedirect.com/science/article/pii/S096585640400103X doi: 10.1016/j.tra.2004.11.001

Bwambale, A., Choudhury, C. F., & Hess, S. (2019). Modelling departure time choice using mobile phone data. *Transportation Research Part A: Policy and Practice*, *130*, 424–439. Retrieved 2021-09-03, from https://www.sciencedirect.com/science/article/pii/S0965856418313569 doi: 10.1016/j.tra.2019.09.054

Bwambale, A., Choudhury, C. F., Hess, S., & Iqbal, M. 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*(48), 2287—2314. Retrieved from https://doi.org/10.1007/s11116-020-10129-5

Börjesson, M. (2008). Joint RP-SP data in a mixed logit analysis of trip timing decisions. *Transportation Research Part E*, *44*(6), 1025–1038. Retrieved from https://doi.org/10.1016/j.tre.2007.11.001

Börjesson, M. (2014). Forecasting demand for high speed rail. *Transportation Research Part A: Policy and Practice*, *70*, 81–92. Retrieved 2023-06-29, from https://www.sciencedirect.com/science/article/pii/S0965856414002444 doi: 10.1016/j.tra.2014.10.010

Börjesson, M., & Eliasson, J. (2014). Experiences from the Swedish Value of Time study. *Transportation Research Part A: Policy and Practice*, *59*(C), 144–158. Retrieved 2023-08-17, from https://econpapers.repec.org/article/eeetransa/v_3a59_3ay_3a2014_3ai_3ac_3ap_3a144-158.htm

Daly, A., & Bradley, M. (1991). Estimation of Logit Choice Models using Mixed Stated Preference and Revealed Preference Information. In *6th International Conference on Travel Behaviour.* Québec.

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.* Retrieved 2020-12-04, from https://trid.trb.org/view/728044 (ISSN: 1474-9122)

Hess, S. (2014). Latent class structures: taste heterogeneity and beyond. In S. Hess & A. Daly (Eds.), *Handbook of Choice Modelling.* Edward Elgar Publishing.

Holmström, A. (2017). *The national travel survey, RVU Sverige (Sweden) 2015-2016* (Tech. Rep.). Stockholm. Retrieved 2022-06-27, from https://www.trafa.se/globalassets/statistik/resvanor/2016/rvu_sverige_2016-reviderad-7-juli.pdf

Holmström, A., & Wiklund, M. (2015). *The national travel survey, RVU Sverige (Sweden) 2011-2014* (Tech. Rep.). Stockholm. Retrieved 2019-09-05, from https://www.trafa.se/globalassets/statistik/resvanor/2009-2015/rvu-sverige-2011-2014.pdf?_t_id=1B2M2Y8AsgTpgAmY7PhCfg%3d%3d&_t_q=rvu+sverige+2011-2014&_t_tags=language%3asv%2csiteid%3af9e4ecf2-4fe2-49ec-bd2f-7b6540d3eb17&_t_ip=62.20.79.10&_t_hit.id=Knowit_EPi_Site_Trafa_KitModules_Document_Models_Media_DocumentFile/_ef44952d-35e8-4716-ab0a-9d20c166d3be&_t_hit.pos=1

Indebetou, L., & Börefelt, A. (2018). *Resvanor invånare 30-49 år i Umeå tätort -Kartläggning med hjälp av ny datainsamlingsmetod hösten 2017* (Tech. Rep.). Umeå. Retrieved 2020-09-08, from https://www.umea.se/download/18.7355677d1631cdbd1241d98/1525342340720/Rapport%20RVU%20Ume%C3%A5%2020170219_v1_1.pdf

Janzen, M. (2019). *Simulating Annual Long-Distance Travel Demand* (PhD Thesis, ETH Zurich). Retrieved from https://doi.org/10.3929/ethz-b-000349744

Kristoffersson, I., Daly, A., Algers, S., & Svalgård-Jarcem, S. (2022). Representing travel cost variation in large-scale models of long-distance passenger transport. *Transportation Letters*, *14*(7), 778–786. Retrieved 2023-11-02, from https://doi.org/10.1080/19427867.2021.1944965 doi: 10.1080/19427867.2021.1944965

Kristoffersson, I., & Liu, C. (2023). Estimation of demand models for long-distance cross-border travel. In *hEART 2023 11th symposium of the European Association for Research in Transportation.* Zurich, Switzerland. Retrieved from https://heart2023.org/abstracts.pdf

McCool, D., Lugtig, P., Mussmann, O., & Schouten, B. (2021). An App-Assisted Travel Survey in Official Statistics: Possibilities and Challenges. *Journal of Official Statistics*, *37*(1), 149–170. Retrieved 2023-06-27, from https://sciendo.com/article/10.2478/jos-2021-0007 doi: 10.2478/jos-2021-0007

Rich, J., & Mabit, S. L. (2012). A Long-Distance Travel Demand Model for Europe. *European Journal of Transport and Infrastructure Research*, *12*(1). Retrieved 2023-06-30, from https://journals.open.tudelft.nl/ejtir/article/view/2946 doi: 10.18757/ejtir.2012.12.1.2946

Stopher, P. R., & Greaves, S. P. (2007). Household travel surveys: Where are we going? *Transportation Research Part A: Policy and Practice*, *41*(5), 367–381. Retrieved from https://doi.org/10.1016/j.tra.2006.09.005

Swedish Transport Administration. (2020). *Analysmetod och samhällsekonomiska kalkylvärden för transportsektorn: ASEK 7.0* (Tech. Rep.).

Varela, J. M. L., Börjesson, M., & Daly, A. (2018). Quantifying errors in travel time and cost by latent variables. *Transportation Research Part B: Methodological*, *117*, 520–541. Retrieved 2023-11-02, from https://www.sciencedirect.com/science/article/pii/S0191261518300870 doi: 10.1016/j.trb.2018.09.010

Wang, X., Shaw, F. A., Mokhtarian, P. L., & Watkins, K. E. (2022). Response willingness in consecutive travel surveys: an investigation based on the National Household Travel Survey using a sample selection model. *Transportation*. Retrieved 2023-04-20, from https://doi.org/10.1007/s11116-022-10312-w doi: 10.1007/s11116-022-10312-w

WSP. (2012). *Höghastighetståg Del 2 – modellutveckling och känslighetsanalyser* (Tech. Rep. No. 2). Stockholm.

Zhang, R., Ye, X., Wang, K., Li, D., & Zhu, J. (2019). Development of commute mode choice model by integrating actively and passively collected travel data. *Sustainability*, *11*(10), 2730. Retrieved from `https://doi.org/10.3390/su11102730`

Appendix A: Descriptive statistics

Table A1: Statistics of mobile phone network dataset variables.

Variable:	**Min:**	**Max:**	**Mean:**	**Median:**	**Std:**
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
Car travel time [min]:	56.44	1136.48	184.65	152.25	102.44
Car cost private [SEK]:	90.09	1685.81	264.85	214.2	151.51
Car cost business [SEK]:	185.19	3465.28	544.42	440.3	311.44
Bus travel time [min]:	35.0	1789.8	323.94	276.95	169.55
Bus number of boardings:	1.0	7.9	2.89	3.0	1.19
Bus first wait time [min]:	2.55	480.0	76.65	53.32	82.7
Bus travel price [SEK]:	74.28	1058.99	196.23	162.11	101.81
Air travel time [min]:	21.6	302.4	118.23	128.56	42.82
Distance to airport [km]:	6.23	382.98	92.85	81.62	47.91
Air first wait time [min]:	12.63	480.0	69.52	48.0	68.37
Air number of boardings:	1.0	3.0	1.69	2.0	0.54
Air travel price [SEK]:	416.85	4581.85	2035.26	2097.24	889.98
Rail travel time private [min]:	11.12	1317.43	191.73	169.73	108.64
Distance to train station [km]:	0.34	502.64	22.14	12.82	28.28
Rail first wait time [min]:	2.82	240.0	35.44	28.25	28.21
Rail number of boardings private:	1.0	5.48	1.95	2.0	0.77
Rail travel price [SEK]:	5.7	2016.55	749.47	665.61	364.24
Rail travel time business [min]:	11.12	1199.57	190.45	168.8	105.12
Rail number of boardings business:	1.0	6.32	2.07	2.0	0.84
Weekend:	0	1	0.29	0.0	0.45
Destination zone density:	0	1	0.02	0.0	0.14

Table A2: Statistics of survey dataset variables.

Variable:	**Min:**	**Max:**	**Mean:**	**Median:**	**Std:**
Car travel time [min]:	57.58	1129.35	197.75	164.32	127.98
Car cost private [SEK]:	46.29	3216.6	443.6	317.31	349.46
Bus travel time [min]:	68.0	1755.45	337.09	286.8	206.11
Bus number of boardings:	1.0	8.7	2.76	3.0	1.13
Bus first wait time [min]:	3.01	480.0	72.99	48.0	80.93
Bus travel price [SEK]:	45.47	1068.83	197.61	167.83	117.14
Air travel time [min]:	21.6	292.36	112.54	123.97	44.49
Distance to airport [km]:	6.01	358.39	94.46	82.3	48.56
Air first wait time [min]:	12.63	480.0	60.64	36.92	71.96
Air number of boardings:	1.0	3.0	1.62	2.0	0.56
Air travel price [SEK]:	416.85	4524.22	1926.82	2003.64	843.79
Rail travel time [min]:	7.21	1211.58	198.04	176.5	132.02
Distance to train station private [km]:	0.72	422.81	28.63	17.24	38.29
Rail first wait time [min]:	2.82	240.0	33.92	25.26	29.52
Rail number of boardings:	1.0	5.62	2.0	2.0	0.81
Rail travel price [SEK]:	5.7	1964.42	752.94	689.79	419.24
Rail number of boardings business:	1.0	6.42	2.12	2.0	0.87
Female solo:	0.0	1.0	0.3	0.0	0.46
Party size:	1.0	420.0	1.99	1.0	5.36
Car competition:	0.0	1.0	0.72	1.0	0.33
No license:	0.0	1.0	0.02	0.0	0.15
Reported business trip:	0.0	1.0	0.12	0.0	0.32
Destination zone density:	0.0	1.0	0.03	0.0	0.17
Weekend:	0.0	1.0	0.26	0.0	0.44

Appendix B: Map of long-distance traffic analysis zones

Figure B1: Map of traffic analysis zones for long-distance trips.
Source: The Swedish Transport Administration.

Appendix C: Definition of regular home/night location

The home antenna of a user is computed by Breyer et al. (2021) 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 the 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 is set to False, otherwise it is True.

Appendix D: Utility functions model II

The utility functions for private (p) for model II trips are

$$V_{i,p,j} = \frac{\phi}{\phi \mathbb{1}_{\{\mathsf{m}\}}(j) + \mathbb{1}_{\{\mathsf{s}\}}(j)} [\kappa_{i,p,j} + \chi_{i,p} t \mathbb{1}_{\{\mathsf{b,a,r}\}}(i) + \beta_p c + \alpha_{i,p} \sigma \mathbb{1}_{\{\mathsf{r}\}}(i) + \omega_p \log(w) \mathbb{1}_{\{\mathsf{f,b,r}\}}(i) + \zeta_p z \mathbb{1}_{\{\mathsf{a}\}}(i)], \tag{9}$$

For business (b) trips:

$$V_{i,b,j} = \frac{\phi}{\phi \mathbb{1}_{\{\mathsf{m}\}}(j) + \mathbb{1}_{\{\mathsf{s}\}}(j)} [\kappa_{i,b,j} + \chi_{i,b} t \mathbb{1}_{\{\mathsf{f,a,r}\}}(i) + \beta_b c + \alpha_{i,b} \sigma \mathbb{1}_{\{\mathsf{f,r}\}}(i) + \omega_b \log(w) \mathbb{1}_{\{\mathsf{f,r}\}}(i) + \eta_{i,b} n \mathbb{1}_{\{\mathsf{f,r}\}}(i)] \tag{10}$$

where $i \in \{\mathsf{f}, \mathsf{b}, \mathsf{a}, \mathsf{r}\}$ denotes the modes air, bus, car or train, data source $j \in \{\mathsf{s}, \mathsf{m}\}$ denotes the data sources survey and mobile phone network data, and $\mathbb{1}_{\{\mathcal{S}\}}(x)$ is the indicator function, which is 1 if $x \in \mathcal{S}$ and 0 otherwise. Furthermore, ϕ is a scaling parameter accounting for differences in the distribution of the error term between the two data sources, κ denotes an alternative specific constant, χ denotes the travel time parameter, t denotes the travel time variable, β is the cost parameter, c is the travel cost, σ is the sum of the distance between the centroid of the origin zone and the closest mode terminal (airport or train station), plus the distance between the centroid of the destination zone and the closest mode terminal, with α as the parameter of the connecting trip representation σ. ω is the parameter belonging to first wait time w (defined as half headway), ζ is the parameter for the dummy variable z, which is 1 if the trip started during a weekend and 0 otherwise and η denotes the parameter for the number of boardings n. The bus alternative is not considered to be available for business trips since business long-distance bus trips are rare, and the alternative specific constant for car is fixed to zero for both trip purposes.

Appendix E: Utility functions model III

The utility functions for private (p) for model III trips are

$$V_{i,p,j} = \frac{\phi}{\phi \mathbb{1}_{\{\mathsf{m}\}}(j) + \mathbb{1}_{\{\mathsf{s}\}}(j)} [\kappa_{i,p,j} + \chi_{i,p} t \mathbb{1}_{\{\mathsf{b,a,r}\}}(i) + \beta_p c + \alpha_{i,p} \sigma \mathbb{1}_{\{\mathsf{r}\}}(i) + \omega_p \log(w) \mathbb{1}_{\{\mathsf{f,b,r}\}}(i) + \zeta_p z \mathbb{1}_{\{\mathsf{a}\}}(i)], \tag{11}$$

and the business (b) utility functions are

$$V_{i,b,j} = \frac{\phi}{\phi \mathbb{1}_{\{\mathsf{m}\}}(j) + \mathbb{1}_{\{\mathsf{s}\}}(j)} [\kappa_{i,b,j} + \chi_{i,b} t \mathbb{1}_{\{\mathsf{f,a,r}\}}(i) + \beta_{b,j} c + \alpha_{i,b} \sigma \mathbb{1}_{\{\mathsf{f,r}\}}(i) + \omega_b \log(w) \mathbb{1}_{\{\mathsf{f,r}\}}(i) + \eta_{i,b} n \mathbb{1}_{\{\mathsf{f,r}\}}(i)] \tag{12}$$

where $i \in \{\mathsf{f}, \mathsf{b}, \mathsf{a}, \mathsf{r}\}$ denotes the modes air, bus, car or train, data source $j \in \{\mathsf{s}, \mathsf{m}\}$ denotes the data sources survey and mobile phone network data, and $\mathbb{1}_{\{\mathcal{S}\}}(x)$ is the indicator function, which is 1 if $x \in \mathcal{S}$ and 0 otherwise. Furthermore, ϕ is a scaling parameter

accounting for differences in the distribution of the error term between the two data sources, κ denotes an alternative specific constant, χ denotes the travel time parameter, t denotes the travel time variable, β is the cost parameter, c is the travel cost, σ is the sum of the distance between the centroid of the origin zone and the closest mode terminal (airport or train station), plus the distance between the centroid of the destination zone and the closest mode terminal, with α as the parameter of the connecting trip representation σ. ω is the parameter belonging to first wait time w (defined as half headway), ζ is the parameter for the dummy variable z, which is 1 if the trip started during a weekend and 0 otherwise and η denotes the parameter for the number of boardings n. The bus alternative is not considered to be available for business trips since business long-distance bus trips are rare, and the alternative specific constant for car is fixed to zero for both trip purposes.
Note that model III also differs from model II in that model III uses information about travel party size and traveller age to compute car and bus travel costs, and the latent class variables also differ (as described in Section 3.1).

Appendix F: Utility functions model IV

The utility functions for private (p) for model IV trips are

$$\begin{aligned} V_{i,p,j} = \frac{\phi}{\phi \mathbb{1}_{\{\mathsf{m}\}}(j) + \mathbb{1}_{\{\mathsf{s}\}}(j)} [&\kappa_{i,p,j} + \chi_{i,p} t \mathbb{1}_{\{\mathsf{b,a,r}\}}(i) + \beta_{p,j} c + \alpha_{i,p} \sigma \mathbb{1}_{\{\mathsf{r}\}}(i) + \\ &\omega_p \log(w) \mathbb{1}_{\{\mathsf{f,b,r}\}}(i) + \zeta_p z \mathbb{1}_{\{\mathsf{a}\}}(i) + \mu g \mathbb{1}_{\{\mathsf{b}\}}(i) \mathbb{1}_{\{s\}}(j) + \iota F \mathbb{1}_{\{\mathsf{b,r}\}}(i) \mathbb{1}_{\{s\}}(j) + \\ &\psi_p y \mathbb{1}_{\{\mathsf{a}\}}(i) \mathbb{1}_{\{s\}}(j) + \lambda_p o \mathbb{1}_{\{\mathsf{a}\}}(i) \mathbb{1}_{\{s\}}(j)], \end{aligned} \tag{13}$$

and the business (b) utility functions are

$$\begin{aligned} V_{i,b,j} = \frac{\phi}{\phi \mathbb{1}_{\{\mathsf{m}\}}(j) + \mathbb{1}_{\{\mathsf{s}\}}(j)} [&\kappa_{i,b,j} + \chi_{i,b} t \mathbb{1}_{\{\mathsf{f,a,r}\}}(i) + \beta_{b,j} c + \\ &\alpha_{i,b} \sigma \mathbb{1}_{\{\mathsf{f,r}\}}(i) + \omega_b \log(w) \mathbb{1}_{\{\mathsf{f,r}\}}(i) + \eta_{i,b} n \mathbb{1}_{\{\mathsf{f,r}\}}(i) + \\ &\psi_b y \mathbb{1}_{\{\mathsf{a}\}}(i) \mathbb{1}_{\{s\}}(j) + \lambda_b o \mathbb{1}_{\{\mathsf{a}\}}(i) \mathbb{1}_{\{s\}}(j)] \end{aligned} \tag{14}$$

where $i \in \{\mathsf{f}, \mathsf{b}, \mathsf{a}, \mathsf{r}\}$ denotes the modes air, bus, car or train, data source $j \in \{\mathsf{s}, \mathsf{m}\}$ denotes the data sources survey and mobile phone network data, and $\mathbb{1}_{\{\mathcal{S}\}}(x)$ is the indicator function, which is 1 if $x \in \mathcal{S}$ and 0 otherwise. Furthermore, ϕ is a scaling parameter accounting for differences in the distribution of the error term between the two data sources, κ denotes an alternative specific constant, χ denotes the travel time parameter, t denotes the travel time variable, β is the cost parameter, c is the travel cost, σ is the sum of the distance between the centroid of the origin zone and the closest mode terminal (airport or train station), plus the distance between the centroid of the destination zone and the closest mode terminal, with α as the parameter of the connecting trip representation σ. ω is the parameter belonging to first wait time w (defined as half headway), ζ is the parameter for the dummy variable z, which is 1 if the trip started during a weekend and 0 otherwise and η denotes the parameter for the number of boardings n, μ is the parameter for travel party size g (which is one if the party size is greater than 8, and zero otherwise), ι is the parameter for female solo traveller F, ψ is the parameter for cars per licence y (restricted to the interval $[0, 1]$ and set to 0 when there are no licences in the household), and λ is the parameter for the complementary dummy variable no licence o. The bus alternative is not considered to be available for business trips since business long-distance bus trips are rare, and the alternative specific constant for car is fixed to zero for both trip purposes.