VTI rapport 1274A

Published 2026

Combining data sources for the estimation of regional travel demand forecasting models

Final report of the FRID project

Ida Kristoffersson, Angelica Andersson, Maria Börjesson, Clas Rydergren och Chengxi Liu

vti

vti.se/publications

VTI rapport 1274A

Combining data sources for the estimation of regional travel demand forecasting models

Final report of the FRID project

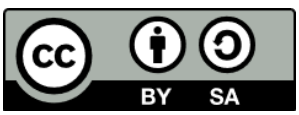
Ida Kristoffersson

Angelica Andersson

Maria Börjesson

Clas Rydergren

Chengxi Liu



Om inget annat anges är publikationen licensierad enligt CC BY-SA 4.0. Omslaget omfattas inte av licensen.

Unless otherwise stated, the publication is licensed under CC BY-SA 4.0. The cover is not included in the license.

vti

Publikationsuppgifter

Titel

Kombinering av datakällor för estimering av prognosmodell för regional transportefterfrågan
Slutrapport i projektet FRID

Författare

Ida Kristoffersson (VTI, orcid.org/0000-0002-3738-9318)

Angelica Andersson (VTI, orcid.org/0000-0002-4926-1434)

Maria Börjesson (VTI, orcid.org/0000-0001-9235-0232)

Clas Rydergren (Linköpings universitet, orcid.org/0000-0001-6405-5914)

Chengxi Liu (VTI, orcid.org/0000-0001-6966-9077)

Utgivare

VTI, Statens väg och transportforskningsinstitut

Serie och nummer

VTI rapport 1274A

Utgivningsår

2026

VTI:s diarienummer

2024/0427/7.1

ISSN

0347–6030

DOI

https://doi.org/10.65151/vti714529

Projektnamn

Framtidens prognosmodeller för regionala resor – att kombinera olika datakällor på bästa sätt (FRID)

Uppdragsgivare

Trafikverket

Publication information

Title

Combining data sources for the estimation of regional travel demand forecasting models
Final report of the FRID project

Author

Ida Kristoffersson (VTI, orcid.org/0000-0002-3738-9318)

Angelica Andersson (VTI, orcid.org/0000-0002-4926-1434)

Maria Börjesson (VTI, orcid.org/0000-0001-9235-0232)

Clas Rydergren (Linköping University, orcid.org/0000-0001-6405-5914)

Chengxi Liu (VTI, orcid.org/0000-0001-6966-9077)

Publisher

VTI (Swedish National Road and Transport Research Institute)

Publication number

VTI rapport 1274A

Published

2026

Registration Number, VTI

2024/0427/7.1

ISSN

0347–6030

DOI

https://doi.org/10.65151/vti714529

Project

Future forecast models of regional travel demand – to combine different data sources in the best way (FRID)

Commissioned by

Swedish Transport Administration

Kort sammanfattning

För robusta beslutsunderlag till investeringar i transportsystemet behövs tillförlitliga beskrivningar av persontrafiken i dagsläget och för ett framtida prognosår. Viktiga komponenter i detta är resornas fördelning på färdmedel, start- och målpunkter i nuläget, samt beteendemodeller för framtida färdmedels- och destinationsval. Tidigare har resvaneundersökningar använts för att ta fram dessa, men kraftigt fallande svarsfrekvenser ger problem med representativitet. Att kombinera resvaneundersökningar med andra passivt insamlade mobilitetsdata ger en möjlighet att förbättra representativiteten.

I denna rapport utvecklar vi metoder för generell datakombination inom två områden: 1) sammanvägd beskrivning av nulägets resmönster och 2) estimering av prognosmodeller för färdmedel- och destinationsval för regionala resor. Vi testar sedan metoderna på en fallstudie för Norrköping med fyra datakällor; resvanedata, mobilnätsdata, smartcard-data från kollektivtrafik, samt GPS-data från fordon. Fallstudien visar att kombination av olika datakällor kan ge en bra bild av nuläget när det gäller resornas fördelning på färdmedel, start- och målpunkter, inklusive att ta hänsyn till datakällornas uppskattade tillförlitlighet.

När det gäller skattning av prognosmodeller för färdmedel- och destinationsval visar fallstudien att det finns flera praktiska utmaningar med att kombinera olika datakällor, i vårt fall framför allt att urvalet av resenärer verkar skilja sig åt mellan datakällorna på ett systematiskt men icke-observerat vis som påverkar valvariabeln. För att samskattning ska ge tillförlitliga resultat behöver vissa antaganden vara uppfyllda för de data-set som modellen skattas på. Vi listar dessa antaganden och diskuterar hur de relaterar till fallstudiens resultat. Sammanfattningsvis visar studien att resvaneundersökningar fortsatt är den datakälla som ger information om reseärende, sällskapsstorlek och socio-ekonomiska data såsom ålder, kön och bilinnehav, men att datakombination har stor potential för att minska beroendet av resvaneundersökningar med låg representativitet. Det är dock fortsatt viktigt att uppmärksamma datakvalitet, urvalets representativitet och de antaganden som ligger till grund för metoderna, om dessa möjligheter ska kunna omsättas i ett tillförlitligt beslutsstöd för transportplanering och investeringsbeslut.

Nyckelord

Datafusion, transportmodell, resvaneundersökning, mobilnätsdata, smartcard data, fordons-gps, efterfrågemodell, prognosmodell, datakombination

Abstract

Robust decision support for investments in the transport system requires reliable descriptions of current travel patterns as well as forecasts for a future target year. Key components include the current distribution of trips across transport modes, trip origins and destinations, and behavioural models for future mode and destination choices. Traditionally, travel surveys have been used to produce such information, but sharply declining response rates have created problems with representativeness. Combining travel survey data with passively collected mobility data offers an opportunity to improve representativeness.

In this report, we develop methods for general data combination in two areas: (1) integrated description of current travel patterns, and (2) forecasting models for mode and destination choice for regional travel. We then test these methods in a case study of Norrköping using four data sources: travel survey data, mobile phone network data, public transport smartcard data, and vehicle GPS-data. The case study demonstrates that combining different data sources can provide a robust representation of current travel patterns, while also accounting for the estimated reliability of the individual data sources.

Furthermore, the case study reveals several practical challenges associated with combining multiple data sources for model estimation. In our case, the most notable issue is that the sample of travellers appear to differ systematically across data sources in ways that are unobserved but also influence the choice variable. For joint estimation to produce reliable results, certain assumptions must hold for the datasets used. We identify these assumptions and discuss how they relate to the findings of the case study. In summary, the study shows that travel surveys remain the main data source for information on trip purpose, party size, and socio-economic characteristics. At the same time, data fusion has considerable potential to reduce reliance on travel surveys with low representativeness. Continued attention must however be paid to data quality, representativeness of the sample and the assumptions underlying the methods if these opportunities are to translate into reliable decision support for transport planning and investment.

Keywords

Data fusion, transport model, travel survey, mobile phone network data, smartcard data, vehicle gps data, demand model, transport forecast, data combination

Table of content

Publikationsuppgifter **4**

Publication information **5**

Kort sammanfattning **6**

Abstract **7**

Sammanfattning **9**

Summary **10**

Preface **11**

1. Introduction **12**

2. Case study Norrköping – Available data **13**

2.1. Survey data 13
2.2. Mobile phone network data 14
2.3. Smartcard data 14
2.4. Vehicle GPS-data 14
2.5. Supply data 14
2.6. Other candidate data sources 15

3. Improved understanding of current travel patterns based on data fusion **16**

3.1. Overview 16
3.2. Method 16
3.3. Implementation details 17

4. Behavioural forecast models based on data combination **19**

4.1. Method 19
4.1.1. Joint estimation 19
4.1.2. Observations sampled from only one choice alternative 20
4.1.3. Inclusion of trip observations lacking mode information 21
4.2. Discussion 22
4.2.1. Necessary assumptions 22
4.2.2. Practical challenges for the forecast choice model 22

5. Conclusions **26**

References **28**

Sammanfattning

Minskande svarsfrekvenser i resvaneundersökningar innebär en växande utmaning för transportplanering och prognosarbete. Resvaneundersökningar utgör fortfarande den viktigaste källan till information om resenärers socioekonomiska förutsättningar och preferenser, men de låga svarsfrekvenserna medför en ökad risk för bristande representativitet. Samtidigt har tillgången till passivt insamlade datakällor, så som mobilnätsdata och smartcarddata, ökat kraftigt under senare år, vilket skapar nya möjligheter att komplettera traditionella datainsamlingsmetoder.

Syftet med detta projekt har varit att undersöka hur flera olika datakällor kan kombineras dels för att förbättra beskrivningen av dagens resande, dels för att användas vid skattning av beteendebaserade färdmedels- och destinationsvalsmodeller. Arbetet har genomförts med Norrköping som fallstudie och baseras på resvaneundersökningsdata, mobilnätsdata, smartcarddata från kollektivtrafiken, automatiska passagerarräkningar samt fordonsbaserad GPS-data.

Projektets första del behandlar datafusion, det vill säga hur information från flera datakällor kan sammanvägas för att ge en mer heltäckande beskrivning av resandet. Inom projektet har en generell och dataagnostisk metod utvecklats för att kombinera olika typer av transportdata. Metoden gör det möjligt att beakta skillnader i datakvalitet, täckningsgrad och representativitet genom att väga samman information från flera källor utifrån deras bedömda tillförlitlighet. Resultaten visar att en kombination av flera datakällor kan ge en mer robust beskrivning av resmönster än vad som är möjligt att uppnå med en enskild datakälla. Metoden möjliggör även användning av datakällor med begränsningar i kvalitet eller representativitet, förutsatt att dessa begränsningar beaktas i analysen.

Projektets andra del behandlar datakombination för beteendebaserade prognosmodeller. Rapporten visar hur etablerade metoder för gemensam modellskattning kan utvidgas till att omfatta flera datakällor som skiljer sig åt avseende information om färdmedel, socioekonomiska variabler och urvalsprocesser. Analysen visar att metoden är teoretiskt genomförbar, men att dess praktiska användbarhet är starkt beroende av kvaliteten hos de ingående datakällorna och av att de antaganden som ligger till grund för skattningen är uppfyllda.

Studiens resultat visar att datafusion utgör en robust och praktiskt tillämpbar metod för att förbättra kunskapen om aktuella resmönster. För beteendebaserade prognosmodeller identifierades däremot flera praktiska utmaningar kopplade till datakvalitet och parameteridentifiering, till exempel att urvalet av resenärer verkar skilja sig åt mellan datakällorna på ett systematiskt men icke-observerat sätt som även påverkar valvariabeln och att det är svårt att skatta en kostnadsparameter för kollektivtrafik när de flesta resor har samma biljettpris. Resultaten understryker särskilt betydelsen av representativa resvaneundersökningar samt vikten av att säkerställa att centrala modellparametrar kan skattas på ett tillförlitligt sätt. Kompletterande datakällor kan bidra till att motverka effekterna av fallande svarsfrekvenser, men kan i nuläget inte ersätta den information som endast finns tillgänglig genom resvaneundersökningar.

En utveckling med stor potential för framtida tillämpningar är den ökande tillgången till högupplösta 5G+-data från mobilnäten. Den förbättrade geografiska upplösningen förväntas ge säkrare information om korta resor och bättre möjligheter att identifiera färdmedel, vilket kan stärka såväl datafusionsmetoder som beteendebaserade prognosmodeller.

Sammantaget visar projektet att kombinationen av flera kompletterande datakällor erbjuder goda möjligheter att förbättra både beskrivningen av dagens resande och underlaget för framtida prognoser. Samtidigt kvarstår behovet av datakällor med tillräcklig kvalitet och representativitet för att resultaten ska kunna användas som ett tillförlitligt underlag för transportplanering, prognosarbete och samhällsekonomiska analyser.

Summary

Declining response rates in travel surveys pose a growing challenge for transport planning and forecasting. While travel surveys remain the primary source of socio-economic information required for behavioural transport models, decreasing representativeness creates uncertainty both in descriptions of current travel patterns and in forecasts used to support infrastructure investment decisions. At the same time, the increasing availability of passively collected data sources, such as mobile phone network data and smartcard data, creates new opportunities to complement traditional survey-based approaches.

This project investigates how multiple transport data sources can be combined both to improve the understanding of current travel demand and to support the estimation of behavioural mode–destination choice models. Using Norrköping as a case study, five data sources were considered: travel survey data, mobile phone network data, public transport smart-card data, automated passenger counts and vehicle GPS data.

The first part of the project develops and demonstrates a generic, data-agnostic data fusion framework for estimating travel patterns by origin, destination and transport mode. The framework combines observations from multiple data sources using a maximum-likelihood approach while explicitly accounting for differences in data quality, representativeness and reliability. A key advantage of the method is that lower-quality data sources can still contribute valuable information without being allowed to dominate the results. The framework therefore provides a practical way of making use of the growing range of mobility datasets available to transport analysts and planners.

The second part of the project investigates how heterogeneous data sources can be used jointly in the estimation of behavioural mode–destination choice models. Building on established methods for combining multiple data sources and handling choice-based sampling, the report presents a framework for jointly estimating models from datasets that differ in terms of available socio-economic information, transport mode information and sampling procedures. The analysis also highlights the assumptions required for such data combination to be valid and discusses practical challenges encountered in the case study.

The results show that data fusion offers a robust and practically applicable approach for improving estimates of current travel demand. The behavioural forecasting framework is theoretically feasible, but its practical performance depends strongly on data quality and representativeness, for example, the sample of travelers appear to differ systematically across data sources in ways that are unobserved but also influence the choice variable, and that it is difficult to estimate the public transport cost parameter when most trips are associated with the same fare. In particular, the case study highlights the continuing importance of representative survey data and the need to avoid identification problems caused by strong correlations between explanatory variables and alternative-specific constants.

Future developments in mobile network data, particularly the increasing availability of high-resolution 5G+ data, are expected to further strengthen both data fusion and behavioural forecasting applications through improved spatial accuracy and potentially better identification of transport modes.

Overall, the project demonstrates that combining multiple complementary data sources is a promising strategy for maintaining and improving transport analysis and forecasting capabilities in the face of declining survey response rates, while also emphasizing that the quality and representativeness of the underlying data remain critical for reliable results.

Preface

This report describes the results and lessons learnt from the FRID-project (Future forecast models of regional travel demand – to combine different data sources in the best way), which has been funded by the Swedish Transport Administration under grant TRV 2024/63017. The project started in October 2024 and ended in October 2026. The project was coordinated by the Swedish National Road and Transport Institute (VTI) with contribution from Linköping University (LiU).

We would like to thank the reviewer Daniel Jonsson (KTH) for constructive feedback at the review seminar conducted online on 2026-10-13, as well as helpful comments from the reference group at the Swedish Transport Administration.

The authors of this report have contributed according to the following CRediT roles:

Ida Kristoffersson: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – original draft

Angelica Andersson: Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft

Maria Börjesson: Conceptualization, Methodology, Supervision, Validation, Writing – original draft

Clas Rydergren: Conceptualization, Data curation, Funding acquisition, Methodology, Supervision, Validation, Writing – original draft

Chengxi Liu: Conceptualization, Methodology, Supervision, Writing – original draft

Stockholm, October 2026

Ida Kristoffersson
Project leader

Granskare/Examiner

Daniel Jonsson, KTH.

De slutsatser och rekommendationer som uttrycks är författarens/författarnas egna och speglar inte nödvändigtvis myndigheten VTI:s uppfattning.

The conclusions and recommendations in the report are those of the author(s) and do not necessarily reflect the views of VTI as a government agency.

Publikationen godkänd för publicering/Publication approved for publication

Magnus Berglund, VTI.

1. Introduction

Traditionally, travel surveys have been used to estimate the parameters of travel demand models used to forecast transport behaviour. However, response rates of such surveys have been declining for decades, to the point where it is no longer reasonable to expect the respondents to be representative of the full population. In Sweden, the response rates to the national travel surveys have gradually dropped from 77,9% in 1994 to 38,2% in 2015 (Trafikanalys 2016), and further to 26% in 2025 (Viklund 2026). Stopher & Greaves (2007) note that travellers with a high valuation of travel time, in other words people who are in a hurry, are less likely to respond to surveys, which would impact travel time valuations calculated from a model based on such a survey. The issue of declining travel survey response rates poses challenges on the local municipality level, as well as in a national and international context.

When social cost benefit analyses are performed, both the forecasted travel behaviour and the current travel behaviour play a crucial role in determining the relative cost-benefit value of different proposed infrastructure projects. The resulting ranking of proposed projects are used to inform investment decisions, which on the national level is managed by Trafikverket and currently amount to roughly 50 billion SEK per year (Forssmed and Carlson 2026).

In this project we consider a case study area of the city of Norrköping, while still discussing our findings in a broader context, as similar challenges with low response rates occur in other geographical areas as well as for long-distance trips. In the two projects Demopan and Demopan2 previously funded by Trafikverket we investigated how passively collected mobile phone network data can be of use for a mode choice model for long distance travel (A. Andersson, Kristoffersson, et al. 2025; A. Andersson, Börjesson, et al. 2025; A. Andersson et al. 2024, 2022). Despite the decreasing response rates, survey data was still found to be a key data source for mode choice modelling, due to the lack of socio-economic and trip purpose information from other data sources. At the same time, the combination with mobile phone network data was found to be beneficial to mitigate the issues related to declining response rates.

In this project we investigate if we can identify a generic method of how to combine several data sources of different types in a mode-destination choice model. To examine the practical feasibility of the proposed method, we use Norrköping as a case study, where we have access to survey data, mobile phone network data, smartcard data and automated passenger counts from the public transport system, and vehicle GPS data. First, we examine how such data sources can be fused to obtain a more accurate description of the current travel situation (Chapter 3). This improved description of the current travel situation is useful for example to evaluate effects of already implemented policies or to investigate accessibility for different user groups. It additionally provides correction terms which are crucial in mode-destination choice model estimation based on combined data sources. Second, we investigate how the same data sources can be used in a joint estimation of a mode-destination choice model, and importantly which assumptions on the used data sources must be fulfilled for such a joint model estimation to be valid (Chapter 4). Throughout the report we also address practical implementation aspects related to the real-world data, which are typical examples of data types which can be accessed or purchased by transport analysts.

2. Case study Norrköping – Available data

The data sources available in the case study represent typical data available to collect or purchase for researchers and practitioners. In this chapter we describe the specific datasets available for the case study as well as other candidate data sources.

The zones illustrated in Figure 1 are the ones considered in the case study. They represent the city of Norrköping as well as the most common commuting areas surrounding the city. A number of data sources are used for the case study: survey data, mobile phone network data, smartcard data including automated passenger counts from the public transport system, and vehicle GPS data. They are described in further detail below.

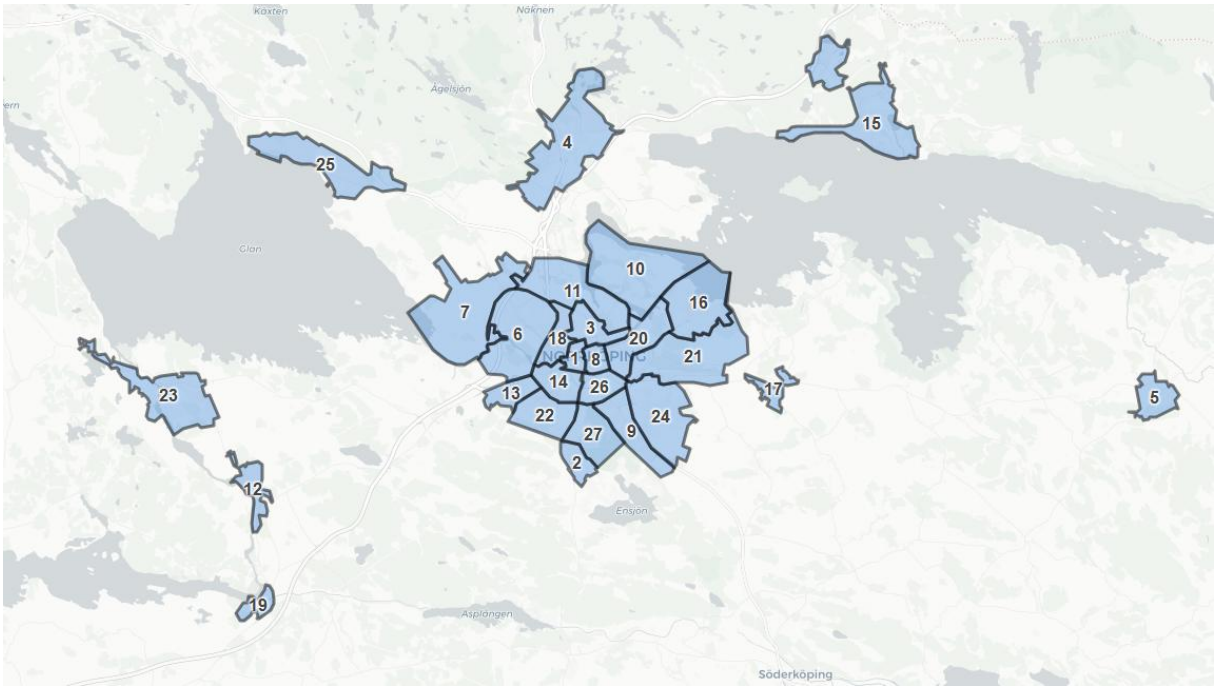


Figure 1: Zone system of Norrköping used in the case study. © Open Street Map.

2.1. Survey data

The survey data comprise 920 trips after data cleaning, and cover ages 16-74. Respondents were asked to report their travels of one measurement day, during a 2-week period in October 2022. The overall response rate was 19%, with younger respondents (<45) represented to a lower degree in comparison to their share of the total population in the studied area. The survey also contains socio-economic information about the respondents including household income, gender, age, children under 16 in household, car access, access to two-wheelers (bike, moped, MC, electric bike), holder of drivers' license, if the respondent has a public transport stop close to home, and if the respondent often or always has a public transport seasonal pass (for instance a monthly or a yearly public transport pass), as well as trip purpose of conducted trips. Note that the survey is the only data source in this project which has information about socio-economic variables and trip purposes, except that smartcard data contains information about if a public transport seasonal pass was used for the trip. There are numerous modes of transport in the survey data, including walk, bike, e-bike, e-scooter, car, public transport, MC and moped. Due to very low response rates of some modes, we consider only the four modes walk, public transport, car and bike (including e-bikes and e-scooters) in this project.

2.2. Mobile phone network data

The mobile phone network data was collected during a 4-week period in 2022 (25/4-22/5). Trips are only represented as number of trips between the zones displayed in Figure 1, without information about transport mode, and also without information about return trips or trip chains. The mobile phone network operator has scaled up the number of trips to represent the full population living in the considered area in a black-box manner, resulting in about 5,8 million trips for the 4-week period. At the time of data collection, the mobile phone network operator had a market share of about 30% on the national level. Municipality representatives suspect that the market share is lower than that in some socio-economically weaker areas, which coincide with areas with lower survey response rates. The data only contain 4G data, which has coverage areas ranging from 200 meters in urban areas to 10 km in sparsely populated areas. This means that it can be difficult to distinguish short trips correctly, especially since the signal can sometimes jump between antennae due to network load, rather than a physical trip by the phone user. The process of filtering out such antenna switches stemming from network load has also been performed by the operator in a black-box manner.

2.3. Smartcard data

The smartcard data cover all tap-ins from the public transport system, roughly 50,000 trips, during a five-day period (2022-09-05 – 2022-09-09). Origin-destination patterns for the public transport trips are inferred by tracking passengers with multiple ticket validations throughout the day to estimate their alighting locations and matching the locations with the considered zone system. These derived trips are then aggregated into daily matrices for each individual public transport line and scaled using an iterative proportional fitting procedure (Knopp and Sinkhorn 1967) to match stop-specific boarding and alighting volumes measured by automated passenger counters equipped on all buses and trams in the studied region. The iterative proportional fitting procedure results in origin stop to destination stop flows consistent with the boarding and alighting data from the automated passenger counts. The resulting data from this combination of tap-ins and automated passenger counts is later in this report referred to as smartcard data. Ticket type is also available in the smartcard data. In the smartcard data, a clear majority of trips are undertaken with some form of a seasonal pass (monthly pass, yearly pass etc.). This contrasts with the survey data, in which a minority of trips by public transport were undertaken by travellers reporting to often or always hold a seasonal pass. This result remains even after weighting the survey responses to match the age and gender distribution in the studied area.

2.4. Vehicle GPS-data

The vehicle GPS data cover about 1.25% of vehicles during a 5-week data collection period in the fall of 2019. The GPS data set was developed within the CODE PROBE project funded by the Swedish Transport Administration (TRV 2021/22404) (Andersson et al., 2024). After removing GPS-trails from trucks, about 3,400 car trips remain. The GPS-trails are mapped to trips between the origin-destination pairs illustrated in Figure 1. Since the GPS data only cover the car mode, the contribution of the GPS-data lies in destination choice for car travel.

2.5. Supply data

Travel times and travel distances per mode were provided by the municipality from their VISUM model. In many cases the VISUM-zones were smaller than the zones used in this project, in such cases we used average values of the smaller zones. All trips by public transport were assumed to have travelled using single tickets, as a ticket choice model was deemed to be outside of the scope of this project. Public transport prices for the year of public transport data collection 2022 were identified from the website of the local public transport provider Östgötatrafiken[1] using WayBackMachine[2]. Public transport ticket

[1] https://www.ostgotatrafiken.se/

[2] https://web.archive.org/

prices range from 27 SEK to 80 SEK, with the vast majority of trips having the price of 27 SEK. Travel cost by car was calculated using a per kilometre price of 1.842 SEK/km[3], this figure is divided by travel party size when available (only for survey trip observations). The municipality also provided destination-specific information per zone, such as shopping areas, recreational areas, number of workplaces and number of school- and kindergarden children admitted within each zone.

2.6. Other candidate data sources

The methods presented in this report are data agnostic and can combine any data source providing origin and destination, and/or mode. Other potential data sources which were not used in this project include data from e-scooter providers, navigation apps, sports apps with or without smart devices such as smart watches, other IoT devices such as helmets with accelerometers and/or GPS data. Also, provided one has a reasonably good route choice model, it would be possible to include traffic count data or traffic flow data for pedestrians, bicyclists, e-scooters and motorized vehicles.

From a transportation perspective, such data sources vary in quality and expected representativeness. For the data fusion method presented in Chapter 3, the inclusion of lower quality data sources is fine, as they can be explicitly set to have a lower reliability than the other included data sources in the optimization scheme. However, for the trade-off's determined by the mode-destination choice model presented in Chapter 4, representativeness is of higher importance for the data combination scheme to be valid. In other words, data sources of too low representativity are better omitted from the choice model.

The mobile phone network data in this project is based on 4G data. High band 5G data, also known as 5G+ data, typically has smaller coverage areas, which from a transportation perspective means that the geospatial resolution of trips is better. With better resolution, the confidence of start- and end zones is higher, and determining transport mode based on mobile phone network data in the case of shorter trips becomes more plausible. 5G+ is quickly becoming more common, especially in urban areas, and should be possible to purchase for any project starting now.

[3] The reimbursement level in 2022 was 1.85 SEK/km (see Inkomstskattelagen 12 kap. 5§, 2007) for combustion engine vehicles, and we have adjusted the cost estimate slightly to account for the (small) share of electric vehicles 2022, which have a lower per kilometre fuel cost.

3. Improved understanding of current travel patterns based on data fusion

To obtain an improved understanding of current travel patterns, we have developed a data agnostic data fusion method. The method is presented in detail in Andersson et al. (2026), so in this chapter we only summarize our contribution and provide some complementary notes on practical implementation which did not fit the WCTR conference paper.

3.1. Overview

In the paper, we develop a framework for combining multiple transport data sources to estimate travel patterns by origin, destination and transport mode. We demonstrate the framework using the case study from Norrköping municipality, where travel survey data, mobile phone network data, public transport smart-card data and vehicle GPS data are used together to produce a more complete picture of travel demand.

The central idea is that each data source captures different aspects of travel behaviour and has its own strengths and weaknesses. By combining these sources in a systematic way, and by incorporating expert knowledge about their reliability, we estimate origin-destination flows that are intended to be more informative than those obtained from any single dataset alone.

The framework is designed to support transport planning and policy analysis by providing improved information on how people travel within a region, while also being flexible enough to be applied in other geographical contexts and with other combinations of data sources.

We address a gap between the growing availability of mobility data and the limited availability of practical methods for combining such data into a coherent description of travel demand. Previous studies have demonstrated the value of data fusion for transport analysis, but most have focused on specific combinations of data sources such as travel surveys and mobile phone data (Brederode et al. 2019; Caceres et al. 2020; Janzen 2019), on estimating selected aspects of travel behaviour such as origin–destination flows or mode-specific demand rather than jointly estimating destinations and modes (Burgalat et al. 2026; Krishnakumari et al. 2019; Tolouei et al. 2017), or on combining data at later stages of transport modelling rather than during data production itself (Andrew Daly and Bradley 1991; Kuhnimhof et al. 2024).

A recurring feature of earlier work is that information on transport mode is often taken directly from a travel survey, while other data sources are primarily used to improve estimates of origin-destination flows. In contrast, our framework aims to estimate both destinations and mode shares using information from multiple data sources simultaneously.

A further contribution is that we explicitly incorporate expert and practitioner assessments of data quality into the fusion process. Rather than assuming that all datasets should be treated equally, we allow local knowledge about representativeness, coverage and reliability to influence the estimation. This helps bridge the gap between methodological development and the realities faced by transport agencies and planners.

More broadly, the paper contributes to the development of data-fusion methods at the data-production stage, where the objective is to create improved travel-demand data before any behavioural modelling is performed. This is an area that has received comparatively less attention despite its increasing importance in a context of declining survey response rates.

3.2. Method

We propose a generic data-fusion framework based on maximum likelihood estimation. The method starts from the assumption that each data source provides an imperfect observation of the same

underlying travel demand. These observations are combined to estimate the most plausible number of trips between origins and destinations and across transport modes.

To enable this, all datasets are first transformed into a common structure describing travel flows between zones. The framework then searches for the set of travel-pattern estimates that best explains the observations across all included data sources simultaneously.

A key feature of the approach is that uncertainty is treated differently for different datasets. We use expert and practitioner knowledge to assess the relative trustworthiness of each source. These assessments determine how strongly each dataset influences the final results.

The framework can accommodate data sources that provide different types of information. For example, some sources may contain information on total travel volumes, while others provide information about specific transport modes. The estimation procedure combines these complementary pieces of information into a unified set of travel-demand estimates.

The proposed methodology is intended to be transferable across study areas and data environments, provided that the available datasets can be translated into a common representation of OD flows and that informed assessments of data quality can be made for the local context.

3.3. Implementation details

This subsection provides a few practical implementation aspects which did not fit the WCTR conference paper (A. Andersson et al. 2026).

In principle, the probability density function (pdf) per data source, which represents the data collection process, can be chosen arbitrarily. In practice however, there are a few things to keep in mind. First, the joint probability function of all observations should preferably have nice convergence properties, which is the case if all data sources are assumed to be represented by normal distributions. Second, it is common that the number inbound trips (zone B → A) per mode are not independent from the number of outbound trips (zone A → B). For all cases where we have pairs of outbound-inbound trips, we must use the bivariate probability density function, rather than just multiplying the two separate pdfs. In other words, the bivariate probability density function needs to be defined for reasonable values of the variances.

One example of when one might want a skewed distribution rather than the normal distribution could be in the case of smartcard data not weighted by automated passenger counts: it is more likely to fail to tap in and travel anyway than to tap in and then not travel.

Regardless of which pdfs are chosen, it is advisable to choose an optimization scheme with an adaptive step size per parameter, to reach convergence within a reasonable time. This is important as the number of parameters to estimate grows quickly with the number of zones and the number of modes: total number of parameters = number of origins * number of destinations * number of modes (if within-zone trips are estimated). For the case study we only considered interzone trips which already with 27 zones and four modes gave us 2,808 parameters to estimate. For the results in Andersson et al. (2026), we use the Adam optimizer, as implemented in Python (v. 3.12.10) Tensorflow (v. 2.20.0). The Adam optimizer is a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments.[4]

Using the Tensorflow library in Python has the additional benefit that it is prepared for GPU accelerated computations. Note that this requires the right hardware, the right installation, and the right implementation in Python. In machine learning contexts, a lot of development has been put into enabling optimization of a large number of parameters using GPU computations for the training of machine learning algorithms such as large language models or image recognition. In the case of our case study, it was possible to run the optimization to convergence using only the CPU, although using the GPU of a

[4] https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/Adam, https://arxiv.org/abs/1412.6980

regular laptop resulted in a speeding up of computational time by a factor of > 100. For a larger number of zones or modes, GPU acceleration is advisable.

It is important to note that the fused results will ultimately depend on the quality of the underlying data. In particular, it is important to consider the coverage per OD-pair and mode. If a particular OD-relation lacks observations altogether, the value of this parameter will not affect the optimization problem, and as such it can result in any arbitrary number due to this underdetermination. Such OD-relations must be assumed to be zero or otherwise removed from the result, to avoid giving the impression of a false level of confidence with respect to the number of trips for that OD-relation. In the case study, there was one origin-destination relation which lacked observations altogether (for all modes and for all data sources), and there were many cases for which there was observations on the total number of trips by all modes (form mobile phone network data) and by some modes, but the split between bike and walk trips was underdetermined. For this reason, the overall plot of the number of trips displayed walk and bike trips added together. If an additional data source, specific to the bike and/or walk mode would have been available, the result would have been more reliable in terms of the walk/bike split. In principle no methodological changes are needed for this, it is only a question of access to data.

4. Behavioural forecast models based on data combination

Behavioural forecasts of future transport demand have traditionally been based on logit models, which in turn are based in utility theory, in which a decision maker makes rational choices aimed at maximizing their own utility. The estimation of such models has traditionally been based on survey data, although in some early cases there has been a need to combine revealed preference surveys, in which the respondent report trips already undertaken, with stated preference surveys, in which the respondent replied what they would have chosen in a hypothetical choice situation. In the following we will assume the reader has knowledge of standard mode-destination choice models, which are described in rich detail by Lerman and Ben-Akiva (1985) among others.

In this chapter we describe how a mode-destination choice model can be estimated based on several data sources. The subsections under 4.1 develop a theoretical framework for combining multiple data sources within a common mode- and destination-choice model when some data sources have socio-economic information while others do not (4.1.1), some data sources are sampled from only one mode of transport while others are sampled from all modes (4.1.2), and some data sources have mode information while others do not (4.1.3). The above subsections show how data-source-specific sampling processes, differences in error variance, choice-based samples, and observations with partially unobserved choices can be accommodated.

The subsections under 4.2 highlight and discuss the assumptions necessary for estimation. The subsection 4.2.1 states the sampling and transferability assumptions required for the parameters to have a common behavioural interpretation across data sources. The subsection 4.2.2 then examines the framework empirically using the data from the Norrköping case study. It considers whether the behavioural parameters can be identified from the variation in these data and whether the assumptions required for a common behavioural interpretation are credible in this application.

4.1. Method

The necessary theory of how to combine two data sources was described already in Daly and Bradley (1991), and of how to estimate a model based on data sampled from a single choice alternative in Manski and Lerman (1977). We extend this in a straight-forward way so that n data sources, rather than just two, can be jointly estimated. The addition of a data source which lack mode information to a mode-destination choice model can be done in two different ways, depending on the empirically determined nesting order of mode and destination. As the nesting order typically varies with trip purpose (see for example Kristoffersson and Liu (2024) for an example where nesting order depends on trip purpose), we describe both options in this report. The most crucial part in determining the feasibility of a joint estimation of a mode-destination choice model lies in the qualitative judgement of whether the available data sources fulfil the assumptions necessary for use of the methods described by Manski and Lerman (1977) and Daly and Bradley (1991), so we discuss this at the end of this subsection.

4.1.1. Joint estimation

The first question to address is how to combine data sources in a joint estimation of a travel demand model. Andersson et al. (2024) illustrate, for long-distance travel mode choice and in line with Daly and Bradley (1991), that two data sources can be jointly estimated, even in cases when some variables are missing from one of the data sources, as long as there is no correlation between the variables only present in one data source and the ones which are present in both data sources. To estimate in this way, a scale parameter ϕ must be estimated and multiplied by the utility functions of one of the two data sources. This accounts for differences in the distribution of the error terms (i.e. differences in variances) between data samples from different data sources. Furthermore, the alternative-specific constants (ASCs) of the different data sources must be kept separate, to account for the possibility that the

probability of being observed using different choices (different modes and/or different destinations) differ between data sources. In both Andersson et al. (2024) and Daly and Bradley (1991), only two data sources are combined. According to Daly and Bradley (1991), the scale parameter ϕ is defined to be the ratio of variances of the two data sources, s_1 and s_2, such that:

$$\phi^2 = \frac{\text{var}(\varepsilon_{s_1})}{\text{var}(\varepsilon_{s_2})} \quad (1)$$

if ϕ is to be multiplied by the utility functions of data source s_2.

Keeping this in mind, the method can be applied to more than two data sources if one data source is left without a scale parameter (as in the case of only two data sources), while all additional data sources have one distinct scale parameter each, formally defined as

$$\phi_{s_2}^2 = \frac{\text{var}(\varepsilon_{s_1})}{\text{var}(\varepsilon_{s_2})} \quad (2)$$

$$\phi_{s_3}^2 = \frac{\text{var}(\varepsilon_{s_1})}{\text{var}(\varepsilon_{s_3})} \quad (3)$$

...

$$\phi_{s_N}^2 = \frac{\text{var}(\varepsilon_{s_1})}{\text{var}(\varepsilon_{s_N})} \quad (4)$$

for data source 2 to *N*, denoted by s_2 to s_N. The utility functions for the joint estimation would then look like:

$$\mathbf{U}_{s_1}(\boldsymbol{\beta}_{s_1}, \boldsymbol{x}_{s_1}) = \mathbf{ASC}_{s_1} + \mathbf{u}_{s_1}(\boldsymbol{\beta}_{s_1}, \boldsymbol{x}_{s_1}) \quad (5)$$

$$\mathbf{U}_{s_2}(\boldsymbol{\beta}_{s_2}, \boldsymbol{x}_{s_2}) = \phi_{s_2}(\mathbf{ASC}_{s_2} + \mathbf{u}_{s_2}(\boldsymbol{\beta}_{s_2}, \boldsymbol{x}_{s_2})) \quad (6)$$

...

$$\mathbf{U}_{s_N}(\boldsymbol{\beta}_{s_N}, \boldsymbol{x}_{s_N}) = \phi_{s_N}(\mathbf{ASC}_{s_N} + \mathbf{u}_{s_N}(\boldsymbol{\beta}_{s_N}, \boldsymbol{x}_{s_N})) \quad (7)$$

where $\mathbf{u}_{s_J}$ denotes the utility function of data source *J* as it would have been specified if it had not been jointly estimated with other data sources, and $\mathbf{U}_{s_J}$ denotes the utility function of data source *J* for the joint estimation, ϕ_{s_J} is the scale parameter of data source *J*, $\mathbf{ASC}_{s_J}$ is the alternative-specific constant for data source *J*, and $\boldsymbol{\beta}$ denotes all other parameters to be estimated and $\boldsymbol{x}$ the independent variables ($\boldsymbol{\beta}$ and $\boldsymbol{x}$ are boldface to indicate that they are matrices containing vectors of parameters/ variables for all choice alternatives). $\mathbf{U}$, $\mathbf{u}$ and $\mathbf{ASC}$ are boldface to indicate that they are vectors containing the specifications of all choice alternatives.

4.1.2. Observations sampled from only one choice alternative

The second question to address is how to make use of data sources that only cover one choice alternative, for example smartcard data, which only cover the mode choice alternative public transport, or vehicle GPS data, which only contain the mode choice alternative car. Zhang et al. (2019) use a correction term along with the regular conditional multinomial logit model, in the context of data combination of active and passive data sources. This correction term was originally derived by Manski and Lerman (1977), and is formulated as part of the estimated alternative-specific constant for the included mode. For example, if data source *K* includes only the choice alternative *m*, then the utility function of data source s_K and choice alternative *m* becomes,

$$\mathrm{U}_{s_K}^m(\boldsymbol{\beta}_{s_K}, \boldsymbol{x}_{s_K}, \phi_{s_K}) = \phi_{s_K} \cdot (ASC_{s_K}^m + u_{s_K}^m(\boldsymbol{\beta}_{s_K}, \boldsymbol{x}_{s_K})) \quad (8)$$

where the alternative-specific constant comprises the two parts

$$ASC_{s_K}^{m} = ASC_{s_K}^{m^*} - ASC_{s_K}^{m'} \quad (9)$$

where $ASC_{s_K}^{m^*}$ is the true ASC, and $ASC_{s_K}^{m'}$ is the correction term, which is based on the ratio between the choice shares in the full population ($Q(m)$) vs in the sample ($H(m)$):

$$ASC_{s_K}^{m'} = \ln(Q(m)/H(m)) \quad (10)$$

In the 1970s the response rates were high so the choice shares of the full population $Q(m)$ could be taken from the result of a survey. Following the contemporary challenges related to declining response rates, we propose instead to take the choice shares of the full population $Q(m)$ from the result of the data fusion method described in Chapter 3.

4.1.3. Inclusion of trip observations lacking mode information

In the case when empirical tests reveal that the destination choice should be placed above the mode choice in the nesting structure, the choice can be set directly to the destination level, similar to Andersson et al. (2025a), in which the choice was set to the road nest level when trip observations lacked information of whether the road trip was by car or by bus. The approach in Andersson et al. (2025a) was inspired by the work done by Daly et al. (2002). This approach allows for an estimation of the mode-unique parameters in the utility function which is in part guided by the other data sources where mode information is available, and in part driven by the transport supply information in the mode specific variables.

The probability of destination d is:

$$P(d) = \frac{exp(W_d + \theta IV_d)}{\sum_{k \in D} exp(W_k + \theta IV_k)}, \quad (11)$$

where

$$IV_d = log\left[\sum_m exp\left(\frac{V_{md}}{\theta}\right)\right] \quad (12)$$

Here W_d is the destination specific utility and V_md is the utility of travelling to destination d by mode m. The logsum IV_d comprises the utility of all modes available to destination d. The chosen mode is not necessary to observe, since the mode utilities are captured by the logsum.

In the opposite case, when empirical tests reveal that the mode choice should be placed above the destination choice in the nesting structure, there are two approaches to handle the situation: 1) variables related specifically to the mode can be omitted for observations for which mode information is missing or 2) the probability of choosing the observed destination can be calculated within each mode nest and probabilities then summed over all modes.

The first approach is valid if there is no correlation between the omitted variables and the remaining ones, analogous with the inclusion or exclusion of socio-economic variables described in the subsection 4.1.1.

In the second approach, the probability of each mode is first calculated as:

$$P(m) = \frac{exp(A_m + \lambda IV_m)}{\sum_j exp(A_j + \lambda IV_j)}, \quad (13)$$

where

$$IV_m = log\left[\sum_{k \in D} exp\left(\frac{V_{mk}}{\lambda}\right)\right] \quad (14)$$

Here A_m is the mode specific utility and IV_m is the logsum over all destination that can be reached by mode m.

The probability of choosing the observed destination, given mode m, is:

$$P(d \mid m) = \frac{exp\left(\frac{V_{md}}{\lambda}\right)}{\sum_{k \in D} exp\left(\frac{V_{mk}}{\lambda}\right)} \quad (15)$$

Since the mode is unknown, the probability of destination d needs to be calculated for each mode m and then summed according to:

$$P(d) = \sum_m P(m)P(d \mid m) \qquad (16)$$

In both nesting structures ('destination above mode' or 'mode above destination'), observations by other data sources which provide information about mode and destination choice are included by setting the choice directly to the observed mode and destination, as usual.

4.2. Discussion

4.2.1. Necessary assumptions

First, the sampling assumption: for each data source, selection into the sample may depend on the chosen alternative and observable strata, but it must not depend systematically on unobserved factors that also affect choice. Differences in observed mode shares arising from choice-based sampling can be addressed using data-source-specific alternative-specific constants, while differences in the variance of unobserved utility can be captured by data-source-specific scale parameters. Neither adjustment can correct for selection related to unobserved preferences.

For example, mobile-network coverage may be systematically lower for trips passing through a particular tunnel. If these trips can be identified through observed origin–destination pairs, the resulting underrepresentation can in principle be addressed through stratification, reweighting or exclusion. This is therefore an example of selection based on observable strata, which is consistent with the sampling assumption, provided that the selection probabilities can be adequately accounted for. In contrast, individuals observed in public-transport data may have a stronger unobserved preference for public transport—and a corresponding dislike of travelling by car—than otherwise comparable individuals with the same socioeconomic characteristics and residential location. If this preference affects both selection into the data source and mode choice, then the selection depends on an unobserved factor that also affects choice. The sampling assumption is then violated, and the resulting selection cannot be corrected for using observed strata or data-source-specific alternative constants.

Second, the transferability assumption: behavioural trade-off parameters in the utility function, such as time and cost sensitivity, must be common across data sources and transferable from the combined sample to the target population. If, for example, the smartcard data and the travel survey represent trips with systematically different time or cost sensitivities, they do not contribute to the same behavioural model. Such differences cannot be absorbed by data-source-specific constants or scale parameters.

For example, smartcard data may contain a larger share of commuting trips than the travel survey. We cannot validate if this is the case since trip purpose is not observed in the smartcard data. If commuting trips have systematically different time or cost sensitivities from other trips, the behavioural trade-off parameters will differ between the data sources. The smartcard data contributes to a behavioural model with larger focus on commuting, compared to the behavioural model estimated on survey data. Data-source-specific constants and scale parameters cannot account for such differences, and the transferability assumption is therefore violated.

4.2.2. Practical challenges for the forecast choice model

In the case study of Norrköping, empirical tests revealed that the mode choice should be placed above the destination choice in the nesting structure. This is consistent with most current Sampers models, although the appropriate nesting structure is an empirical matter and may differ across models and trip purposes. In the long-distance Sampers model, mode choice is placed above destination choice for private and commuting trips, whereas the reverse structure is used for business trips. In the regional Sampers models, mode choice is placed above destination choice for commuting, recreation, grocery shopping, other shopping, compulsory-school, visiting, home-based business, work-based business, escort, service/childcare/healthcare, and other trips. The reverse structure was initially used for upper-

secondary-school and adult-education trips. However, the logsum parameter was ultimately fixed at one, causing the nested logit structure to collapse to a multinomial logit model (Kristoffersson et al. 2024).

In the Norrköping case study of this project, all trip purposes were modelled together as there are few observations in the survey data, which is the only data source with information about trip purpose. Following the empirical result of the nesting structure, mobile phone network trip observations, which lack mode information, were included by omitting mode-specific variables for mobile phone network data observations after verifying that the variables were not correlated, i.e. according to the approach 1) described in subsection 4.1.3. This approach was chosen for the Norrköping case study since it was substantially easier to implement in the Biogeme estimation software. Vehicle GPS data and smartcard data were included in the estimation with a correction term for the alternative-specific constant (ASC), according to the theory described in the subsection 4.1.2. Also, the estimation was performed as described in subsection 4.1.1 for all data sources.

The correction for the choice-based sampling of the smartcard data relies on the mode shares estimated by the data-fusion model in Chapter 3. The behavioural model is therefore conditional on the results of this first estimation stage. Uncertainty or systematic errors in the estimated mode shares may be carried forward through the correction terms and affect the behavioural parameter estimates. Moreover, even correctly estimated mode shares cannot correct for selection related to unobserved preferences. The sensitivity of the results to alternative plausible mode shares may therefore be examined in further studies.

The estimation process performed on the data from the case study revealed several practical problems, that limit the usefulness of the specific estimation results. Still, highlighting the practical challenges that occurred has relevance also for other contexts.

In order to confirm that the data preparation and model implementation worked as intended, the model was first estimated using the travel-survey data alone, after which the other data sources were added sequentially. When the smartcard data were added, using weights to correct for their choice-based sampling and reproduce the estimated public-transport mode share in the target population, the estimated cost parameter became positive. Such an estimate has no plausible behavioural interpretation, as it would imply that travellers prefer higher travel costs, and therefore indicates that the common cost parameter cannot be estimated reliably from the available data.

Several factors may contribute to this result. First, public-transport fares exhibit limited variation in the case-study data. Most observed trips take place within Norrköping and are subject to the same fare, while relatively few trips cross into surrounding fare zones. This makes it difficult to distinguish the effect of cost from the data-source-specific alternative-specific constant for public transport, as reflected in the correlation exceeding 0.9 between their parameter estimates. This is primarily a limitation of the empirical application rather than of the general method: greater independent fare variation in another application could improve identification of the cost parameter. Removing either the public-transport constant or the cost parameter would not provide a satisfactory solution. The constant is required to accommodate systematic differences in public-transport shares across data sources, while removing the cost parameter would prevent the model from estimating a central behavioural trade-off.

Second, the data sources may represent different types of travellers or trips. For example, the smartcard data lack information on car access, driving-licence holding and trip purpose. If the travellers or trips represented in these data differ systematically from those in the travel survey in ways that also affect cost sensitivity, the assumption of a common cost parameter across data sources may be violated. Such differences cannot necessarily be addressed by weighting or data-source-specific constants, particularly when they are related to unobserved preferences. Other differences in variable availability or measurement across the data sources may also contribute to the estimation problem.

The present analysis cannot determine whether the positive estimate is primarily caused by weak identification due to limited fare variation, selective representation of travellers and trips, omitted information in the smartcard data, or a combination of these factors. More generally, the case study illustrates a fundamental challenge in using smartcard data to estimate common behavioural parameters

jointly with other data sources. Smartcard observations constitute a non-random, choice-based sample of public-transport users and trips. These travellers may differ from the target population not only in observed characteristics but also in unobserved factors, such as a stronger preference for public transport or a greater aversion to car travel. If such unobserved factors affect both inclusion in the smartcard data and mode choice, weighting based on observed characteristics or estimated mode shares cannot fully correct the selection. In the present case study, this potential selection problem, together with the limited fare variation, prevents reliable estimation of the common cost parameter. Whether these problems can be overcome in other applications depends on both sufficient fare variation and the availability of information or modelling assumptions that can account for selection into the smartcard data.

In Figure 2 we have highlighted the key components and assumptions that need to be in place for a joint estimation of a choice model to be valid. Keeping these in mind, consider how these key aspects relate to the occurrence of a cost parameter which is highly correlated with the ASC. Due to the high correlation between the cost parameter and the ASC for public transport, point 1 in Figure 2: that ASC should be separated per data source, is challenging to achieve as some of the effect of the ASC risk ending up in the cost parameter, which should optimally be shared between modes according to point 4. As the smartcard data is only available for one mode, this also means that the cost parameter needs to be shared between data sources.

Required prerequisites of the data samples:

The sampling assumption: For each data source, selection into the sample may depend on the chosen alternative and observable strata, but it must not depend systematically on unobserved factors that also affect choice.

The transferability assumption: Behavioural trade-off parameters in the utility function, such as time and cost sensitivity, must be common across data sources and transferable from the combined sample to the target population.

Important requirements for model specification:

1. ASCs should be separated per data source – to account for the possibility that the probability of being sampled by a certain mode differ between data sources.
2. A data-source specific scale parameter ϕ should be multiplied to the utility functions – to account for the possibility of differing noise levels between data sources.
3. Data sources that are sampled from only one choice alternative (mode) should include an ASC correction term – to account for the share of that mode among all modes.
4. The cost parameter should be the same for all modes – to account for that marginal utility of money should remain constant within one individual. Using a shared cost parameter also has the positive side effect of reducing potential problems with lack of trade-off situations in the data per mode.

Figure 2. Summary of key prerequisites for joint estimation of a mode-destination choice model using several data sources.

Furthermore, in the case of smartcard data, which is sampled from only one mode, a correction term for the ASC is also needed according to point 3 in Figure 2. As such, in case of any misspecification of the true mode shares, which form the basis of the correction term, there is a risk that such a misspecification could propagate through the estimated ASC to affect the cost parameter.

One of the main reasons to combine several data sources is to obtain more trade-off situations in the data material, so that the parameter values can be estimated more accurately. In principle, as long as all requirements listed in Figure 2 are fulfilled, one data source should not be able to "drown out" other data sources based on just having a larger number of observations. However, as discussed above, point 1 is likely not fulfilled in the data material from the case study. Furthermore, the notable difference in overall ticket choices between survey data and smartcard data discussed in Chapter 2 gives rise to some concern. This could indicate that the sampling assumption in Figure 2 is also not fulfilled. The uncertainty in the fulfilment of these two requirements poses a risk that the significant changes that occur in the cost parameter when the smartcard data is added represents the smartcard data drowning out information from other data sources, rather than enhancing the quality of the estimation of the cost parameter.

If the smartcard data do not comply with the sampling assumption of Figure 2, they should not be used. However, if this assumption is not violated, one option would be to exclude the cost parameter from the public transport utility function altogether and allow its effect to be captured by the ASC. However, this would probably not produce useful model output, since the cost parameter would then affect only the utility function for car. Moreover, car cost is highly correlated with travel time because it is calculated using a per-kilometre rate. Andersson et al. (2024) and Andersson et al. (2025b) noted in the case of mobile phone network data, that the use of a per-kilometre price in the absence of information about party size resulted in a lower quality of estimation results in terms of the trade-off between travel time and travel cost. In the case study in this report, there are large numbers of trip observations by both vehicle GPS data and mobile phone network data which lack complementary information about car cost, such as party size and fuel type. It is also relevant to note that such uncertain identification of travel cost risk causing attenuation bias in the estimate of the cost parameter. In the presence of attenuation bias, the variable noise causes the parameter estimate to be lower than what it should be, which in the case of the cost parameter also leads to an over-estimation of the inferred value of travel time.

Due to the long array of issues related to the data quality of the case study, we choose to not present the estimation results, to make sure that they are not accidentally used when they are in fact too uncertain for practical use.

Even though the method presented in this report provides a feasible way forward in the estimation of a mode-destination choice model on several data sources, the practical applicability of the method ultimately depends on a minimum level of data quality of the included data sources. As has been stated before (A. Andersson et al. 2024), the introduction of novel data sources does not automatically mean that travel survey data can be replaced, only that the quality of the estimation can be maintained even in the face of declining response rates. At this point in time, travel survey is still the main source for information about travel party size, age, sex, drivers' license, trip purpose and car access that are important for accurate travel demand forecast models. As stated earlier in this report, this requires that the survey sample is reasonably representative of the full population. In other words, we are faced with the curse of declining response rates: when response rates fall, the need for data combination increases, but when response rates fall, the risk that the survey does not meet the minimal requirements for valid data combination also increases.

It may be possible that some socio-economic information could be provided via other data sources. For instance, some mobile phone operators offer to share information about age and sex as well as trip information. It is important to stress though that the data source that also provide socio-economic information should preferably provide mode information and include all modes of transport. If socio-economic information is shared without the active consent of the traveller, it is important to carefully consider respect of integrity and compliance with integrity laws. If instead socio-economic information is shared with the active consent of the traveller, careful consideration of whether there is risk of selection bias in the sample is needed.

5. Conclusions

This project has investigated how multiple transport data sources can be combined both to improve the description of current travel patterns and to support the estimation of behavioural forecast models. Using Norrköping as a case study, we have demonstrated both the opportunities and the practical challenges associated with combining traditional travel survey data with different forms of passively collected mobility data.

A key contribution of the project is the development and demonstration of a generic, data-agnostic data fusion framework capable of combining heterogeneous transport datasets into a unified representation of travel demand. Unlike many previous applications that focus primarily on improving origin–destination estimates, the proposed framework enables the simultaneous estimation of both origin–destination flows and mode shares.

An important advantage of the framework is its ability to explicitly account for differences in data quality. By incorporating expert assessments of representativeness, coverage and reliability, data sources can be weighted according to their expected trustworthiness. This means that data sources of lower quality do not necessarily need to be excluded from the analysis. Instead, they can contribute information while exerting a proportionally smaller influence on the fused result. Consequently, the framework provides a practical way of making use of the increasingly diverse mobility datasets available to transport planners and analysts.

At the same time, the quality of the fused output remains dependent on the quality and coverage of the underlying data. Data fusion cannot compensate for a complete absence of observations, and insufficient information for specific origin–destination–mode combinations may still result in underdetermined estimates. Continued efforts to improve both data coverage and representativeness therefore remain important. This is especially true for estimating the split between walking and cycling.

The project has also shown that it is theoretically possible to estimate mode–destination choice models using a combination of several data sources with different characteristics. Existing methods for combining revealed preference datasets, handling choice-based sampling, and incorporating observations lacking information on all choice dimensions has within the project been extended in a consistent manner to multiple data sources. However, data combination for estimation of mode-destination choice models imposes two key requirements on each data source. First, selection into the sample may depend on the chosen alternative and observable strata, but it must not depend systematically on unobserved factors that also affect choice. Second, behavioural trade-off parameters in the utility function, such as time and cost sensitivity, must be common across data sources and transferable from the combined sample to the target population.

The case study also highlighted the complementary strengths and limitations of the individual data sources. The travel survey provided essential information on trip purposes, party size and socio-economic characteristics, but suffered from a low response rate and underrepresentation of younger respondents. Mobile-phone network data offered large-scale origin–destination coverage but had limited spatial accuracy for short trips and did not distinguish between modes. Smartcard data provided detailed information on public-transport trips but covered only one mode and differed substantially from the travel survey in terms of observed ticket choices. These differences raise concerns about whether the sampling and transferability assumptions stated above hold for the smart-card data. Vehicle GPS data provided extensive information on observed car trips and their destination choices, but missing information on party size and fuel type made the calculation of car costs uncertain, and it had no information about other modes than car.

The case study also revealed an identification problem. In the Norrköping case, the rather small area resulted in limited variation in trip lengths, car travel times and costs, as well as public-transport travel times and fares. Particularly when adding the smart-card data there was a high correlation between the public-transport cost parameter and the alternative-specific constant. This made it difficult to identify cost sensitivity separately from the average attractiveness of public transport, complicated the joint

estimation and risked producing implausible behavioural parameters. Furthermore, the size of zones may be problematic for identification of walk and bike trips if a large share of these trips are within-zone trips. Future applications should therefore assess the plausibility of the selection and transferability assumptions and ensure that the data and modelled area provide sufficient independent variation to identify the key behavioural parameters.

More broadly, the results reinforce the importance of obtaining representative survey data. While additional data sources can complement travel surveys and help mitigate declining response rates, they cannot completely replace the information currently available only through surveys, such as trip purpose, travel party composition, car availability, licence holding and other socio-economic characteristics such as age, gender and income. Since the validity of data combination methods depends on the transferability of behavioural trade-offs from the observed sample to the wider population, representativeness remains a fundamental requirement.

As for the data fusion framework, future access to 5G+ mobile network data may improve behavioural forecasting applications. Better spatial resolution and potentially more reliable mode identification could strengthen both the estimation of behavioural parameters and the correction terms used when combining multiple datasets.

Overall, the findings suggest that the future of transport demand analysis is likely to rely increasingly on the combination of multiple complementary data sources rather than any single source alone. Data fusion offers a practical and immediately applicable way of improving the description of current travel patterns, while data combination for behavioural estimation provides a promising path towards more robust forecast models. However, separating data fusion and behavioural estimation into two stages implies that uncertainty within the data fusion stage is not accounted for in the behavioural estimation stage. Future work could therefore improve the modelling by conducting data fusion and estimation of mode-destination choice models in a joint manner.

References

Andersson, Angelica, Maria Börjesson, Nils Breyer, Andrew Daly, Leonid Engelson, and Ida Kristoffersson. 2025. 'Estimating a Nested Latent Class Mode Choice Model on Mobile Network Data'. *Transportation Planning and Technology* 0 (0): 1–33. https://doi.org/10.1080/03081060.2025.2593435

Andersson, Angelica, Leonid Engelson, Maria Börjesson, Andrew Daly, and Ida Kristoffersson. 2022. 'Long-Distance Mode Choice Model Estimation Using Mobile Phone Network Data'. *Journal of Choice Modelling* 42 (March): 100337. https://doi.org/10.1016/j.jocm.2021.100337

Andersson, Angelica, Ida Kristoffersson, Andrew Daly, and Maria Börjesson. 2024. 'Long-Distance Mode Choice Estimation on Joint Travel Survey and Mobile Phone Network Data'. *Transportation Research Part A: Policy and Practice* 190 (December): 104293. https://doi.org/10.1016/j.tra.2024.104293

Andersson, Angelica, Ida Kristoffersson, Clas Rydergren, and Maria Börjesson. 2025. 'Modelling Travel Mode Choice on Combined Data Sources Using Machine Learning Methods'. *Transportation Research Procedia*, VSI: TRPRO_EWGT 2024, vol. 86 (January): 88–95. https://doi.org/10.1016/j.trpro.2025.04.012

Andersson, Angelica, Chengxi Liu, Clas Rydergren, Ida Kristofferson, and Maria Börjesson. 2026. 'Data Fusion for Travel Pattern Estimation'. Paper presented at World Conference of Transport Research. *Proceedings of WCTR 2026*.

Andersson, Malin, Joakim Ahlberg, Marian Codreanu, et al. 2024. *CODE PROBE - Continuous Travel Demand and Link Flow Estimation Based on GPS Data*. Linköpings Universitet, Ramboll. https://urn.kb.se/resolve?urn=urn:nbn:se:trafikverket:diva-19968

Brederode, Luuk, Mark Pots, Ruben Fransen, and Jan-Tino Brethouwer. 2019. 'Big Data Fusion and Parametrization for Strategic Transport Demand Models'. *2019 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, June, 1–8. https://doi.org/10.1109/MTITS.2019.8883333

Burgalat, Joan, Gael Pallares, Myriam Foucras, and Yohan Dupuis. 2026. 'A Literature Review of Public Transport OD Matrix Estimation'. *Future Transportation* 6 (1): 45. https://doi.org/10.3390/futuretransp6010045

Caceres, Noelia, L. M. Romero, and Francisco G. Benitez. 2020. 'Exploring Strengths and Weaknesses of Mobility Inference from Mobile Phone Data vs. Travel Surveys'. *Transportmetrica A: Transport Science* 16 (3): 574–601. https://doi.org/10.1080/23249935.2020.1720857

Daly, A., J. Fox, and C. Rohr. 2002. 'Advanced Modeling to Overcome Data Limitations in the Norwegian Transport Model'. Paper presented at European Transport Conference 2002MVA, Limited; Association for European Transport. *Publication of: Association for European Transport*. https://trid.trb.org/view/728044

Daly, Andrew, and Mark Bradley. 1991. 'Estimation of Logit Choice Models Using Mixed Stated Preference and Revealed Preference Information'. Paper presented at 6th International Conference on Travel Behaviour, Québec. *6th International Conference on Travel Behaviour*.

Forssmed, Jakob, and Andreas Carlson. 2026. *Nationell Planering För Transportinfrastrukturen 2026–2037*. Regeringens Skrivelse 2025/26:259. Stockholm. https://www.regeringen.se/rattsliga-dokument/skrivelse/2026/04/skr.-202526259

Janzen, Maxim. 2019. 'Simulating Annual Long-Distance Travel Demand'. PhD Thesis, ETH Zurich. https://doi.org/10.3929/ethz-b-000349744

Knopp, P., and R. Sinkhorn. 1967. *Concerning Nonnegative Matrices and Doubly Stochastic Matrices*. Pacific J. Math., 21(2): 343–48. http://dx.doi.org/10.2140/pjm.1967.21.343

Krishnakumari, Panchamy, Hans van Lint, Tamara Djukic, and Oded Cats. 2019. 'A Data Driven Method for OD Matrix Estimation'. *Transportation Research Procedia*, Journal of Transportation and Traffic Theory, vol. 38 (January): 139–59. https://doi.org/10.1016/j.trpro.2019.05.009

Kristoffersson, Ida, Svante Berglund, Sandra Samuelsson, Peter Almström, and Staffan Algers. 2024. *Sampers 4 Skattning Av Regionala Efterfrågemodeller Implementationsversion*. https://bransch.trafikverket.se/contentassets/ef39a7c8e2964def82ee114d4d3798b5/2024/sampers-4---skattning-av-regionala-efterfragemodeller-implementationsdokumentation-2024-04-02.pdf

Kristoffersson, Ida, and Chengxi Liu. 2024. 'Estimation of Demand Models for Long-Distance International Travel–Key Determinants for Swedes' Travel Abroad'. *European Journal of Transport and Infrastructure Research* 24 (4): 62–88. https://doi.org/10.59490/ejtir.2024.24.4.6696

Kuhnimhof, Tobias, Léa Fabre, and Mario Cools. 2024. 'Workshop Synthesis: Data Fusion - Generating More Than a Sum of Parts'. *Transportation Research Procedia*, 12th International Conference on Transport Survey Methods, vol. 76 (January): 670–77. https://doi.org/10.1016/j.trpro.2023.12.090

Manski, Charles F., and Steven R. Lerman. 1977. 'The Estimation of Choice Probabilities from Choice Based Samples'. *Econometrica: Journal of the Econometric Society*, 1977–88. https://doi.org/10.2307/1914121

Stopher, Peter R., and Stephen P. Greaves. 2007. 'Household Travel Surveys: Where Are We Going?' *Transportation Research Part A: Policy and Practice* 41 (5): 367–81. https://doi.org/10.1016/j.tra.2006.09.005

Tolouei, Reza, Stefanos Psarras, and Rawle Prince. 2017. 'Origin-Destination Trip Matrix Development: Conventional Methods versus Mobile Phone Data'. *Transportation Research Procedia*, Emerging technologies and models for transport and mobility, vol. 26 (January): 39–52. https://doi.org/10.1016/j.trpro.2017.07.007

Trafikanalys. 2016. *Resvaneundersökningar Som Indata till Persontransportmodeller - Problem, Möjligheter Och Framtida Behov i Sverige Och Norge*. 2016:21. http://www.trafa.se/globalassets/rapporter/rapport-2016_21-resvaneundersokningar-som-indata-till-persontransportmodeller--problem-mojligheter-och-framtida-behov-i-sverige-och-norge.pdf

Viklund, Mattias. 2026. *Resvanor i Sverige 2025 - Kvalitetsdeklaration*. Trafikanalys. https://www.trafa.se/globalassets/statistik/resvanor/2026/kvalitetsdeklaration-resvanor-i-sverige-2025.pdf

Zhang, Ruone, Xin Ye, Ke Wang, Dongjin Li, and Jiayu Zhu. 2019. 'Development of Commute Mode Choice Model by Integrating Actively and Passively Collected Travel Data'. *Sustainability* 11 (10): 2730. https://doi.org/10.3390/su11102730

VTI (Swedish National Road and Transport Research Institute) is Sweden's independent institute for applied transport research. On behalf of the Swedish Government, we conduct research and development that addresses real societal needs and contributes to an efficient, safe and sustainable transport system.

Our research covers all modes of transport and ranges from people, vehicles and infrastructure to planning, economics and policy instruments. Within a multidisciplinary organisation, we combine specialist expertise with a broad systems perspective and advanced research environments. Through a unique research infrastructure, we produce independent evidence and decision support used in both national and international transport policy.

VTI works in close collaboration with academia, public authorities and industry. We are around 250 people working together from offices in Linköping, Stockholm, Gothenburg, Lund and Skellefteå.

vti