Combined mobile network data and survey data to describe travel patterns – method memorandum

PM: 2025:8

Date: 2025-05-21

TRANSPORT ANALYSIS

Transport Analysis
Adress: Rosenlundsgatan 54,
SE-118 63 Stockholm
Phone: +46 10 414 42 00
E-mail: trafikanalys@trafa.se
Webadress: www.trafa.se
Publisher: Mattias Viklund
Datum: 2025-05-21

Foreword

Transport Analysis describes the daily travel habits of the Swedish population through an annual travel survey based on questionnaire responses.

In 2023, Transport Analysis conducted an analysis of how mobile network data can be used to describe travel patterns. We concluded that it is possible to break down mobile network data by month and by county in order to describe travel volumes. At the same time, we identified shortcomings in using mobile network data to describe travel relations. The next step was to identify appropriate methods for combining mobile network data with the existing travel survey.

This memorandum presents analysis and methodological work made by Transport Analysis on combining questionnaire data from the annual travel survey with mobile network data, to describe travel patterns.

Björn Tano has been the project manager. Andreas Holmström and Filippa Egnér were members of the project team. Heads of Division Sofie Orrling and Andreas Tapani participated in the steering group.

Stockholm, May 2025

Sofie Orrling

Head of Division

Table of Contents

1 Summary

Transport Analysis describes the daily travel behaviour of the Swedish population using a well-established questionnaire-based survey: the national travel survey *Resvanor i Sverige*, also referred to as RVU[1].

As response rates in survey-based studies decline, access to alternative data sources is increasing. It is therefore important to keep pace with this development. In this project, we implemented methods to combine mobile network data with RVU data for the survey years 2020–2023. The aim was to combine the two sources – mobile network data and RVU – in a way that ensures high statistical quality. This memorandum presents the methodological work behind this implementation.

In a previous analysis[2], we concluded that mobile network data can be broken down by month and by county (within-county travel). Based on these two levels of aggregation, we developed and tested methods to identify the most suitable solution for estimating travel volumes.

Several different data sources and combinations of sources are available for estimating travel volumes:

a) Mobile network data
b) RVU data
c) A combination of both sources
d) Other data sources

Our starting point was to apply method (c), that is, a combination of mobile network data and RVU data. This memorandum describes how we analysed and tested different approaches to identify the most suitable one. The results showed that a robust combined model – integrating a linear mixed model with transfer learning – was the most suitable for estimating the number of trips per month using the two data sources. At the county level, a pure transfer learning method was found to be the most appropriate for estimating travel volumes.

It is desirable to use a robust combined model, and we found that this is possible at the monthly level. However, at the county level we observed outliers that heavily affect linear models, referred to as influencers. In such cases, linear prediction becomes risky. Transfer learning is more resistant to these influencers than linear models.

The results from this project are not part of Sweden's official statistics[3] and are published on Transport Analysis website: www.trafa.se/transportmonster/RVU-Sverige/kombinerade-mobilnatsdata-och-enkatdata-beskriver-resmonster-15129. The publication includes both interactive tables and figures.

[1] Trafikanalys, "Resvanor, Sverige"
[2] Trafikanalys, ”PM 2023:6 Hur väl kan mobilnätsdata beskriva våra resvanor?”
[3] SCB, ”Sveriges officiella statistik”

2 Introduction

Knowledge about how people in Sweden travel is an important basis for designing a future transport system that can sustainably meet both current and future societal challenges. There is therefore a need to develop methods for capturing travel patterns in a cost-effective way, based on the best available data sources.

Traditionally, travel surveys have been used for this purpose. However, as most individuals now carry a mobile phone, mobile network data – that is, information about which base stations users are connected to – has been proposed as a complementary source to today's travel surveys. The current methods used in travel surveys also face certain challenges, including limitations in coverage, high costs and the burden placed on respondents. There is also a growing need to capture more detailed and comprehensive travel relations, which mobile network data may help to address.

In 2023, we analysed[4] whether and how Telia's mobile network data could be used to describe the travel patterns of the Swedish population. The conclusions were twofold: we identified some issues with the reliability of Telia's data, but also saw potential in using mobile network data. However, such data lacks information about the purpose of the trip, mode of transport and who is making the journey – information that is a crucial part of travel surveys.

2.1 Purpose and objectives

As a next step, we proposed investigating how mobile network data could be combined with the national travel survey, RVU[5]. Based on the conclusions from that project, Transport Analysis carried out further work to explore how such an implementation could be realised. This memorandum presents the findings from that implementation work, with the aim of developing the methods we use to capture travel behaviour.

The objective was to identify robust methods that can be used both to describe travel patterns and to serve in other contexts where mobile network data is available in addition to survey data. A broader aim is, in other words, to keep pace with developments and use mobile network data in combination with survey data in other areas as well. However, this does not mean that traditional surveys should be discarded.

In the best case, mobile network data may be used in combination with survey data as a starting point. In the future, it may become relevant to rely entirely on mobile network data, but we are not there yet. The focus of this project has been to identify and describe robust methods for combining mobile network data with survey data from the travel survey.

2.2 Structure of the memorandum

This memorandum describes methods for combining the two data sources: mobile network data and RVU data. In Chapter 3, we present the magnitude of the data sources in terms of

[4] Trafikanalys, ”PM 2023:6 Hur väl kan mobilnätsdata beskriva våra resvanor?”
[5] Trafikanalys, ”Resvanor, Sverige”

the number of trips, as well as the correlation between them. The methods are then described in Chapter 4.

In Chapter 5, we report on the analyses conducted and the methodological choices made based on these, covering trips per month, within-county trips, and trips by mode and purpose.

Finally, we present conclusions and a brief discussion in Chapter 6, outline new publication activities in Chapter 7, and discuss future opportunities in Chapter 8.

3 Data and conditions

In a previous project by Transport Analysis, we concluded that there was a good correlation between mobile network data and RVU data for the years 2019–2021. However, there were significant differences in the total number of trips reported. Since then, Telia has made methodological changes in how they weight up the number of trips. See Telia's response[6] regarding why their data has changed:

"The new version is more stable than the previous one and is less affected by random fluctuations in signal quality. It also includes improvements in our ability to extrapolate from our users to the total population – a development that has taken place in cooperation with Statistics Sweden (SCB). Additionally, we have expanded the data set to include corporate subscriptions from 2020 onwards. Overall, one of the outcomes is that the figures are lower than in our previous version, V1.

The data is comparable year to year, but in 2019 we did not include corporate subscriptions, which may mean that the number of trips is somewhat underestimated."

Due to Telia's revised calculation algorithm, the level differences between the two data sources have decreased. The total number of trips is lower in the new version of Telia's data compared to the previous version. Because of the uncertainty surrounding the 2019 mobile network data, we have chosen to exclude the 2019 data and instead base our analyses in this memorandum on data from the years 2020–2023.

Telia's mobile network data includes a variable describing signal quality, and Telia has assessed that the best estimate of the total number of trips is obtained by using the entire dataset, regardless of signal quality. In this memorandum, we have included all signal data from Telia, in accordance with their recommendation from our earlier analysis.

3.1 Data

See Figure 1 for a summary of the number of trips by data source. For the years 2020–2023, the development differs between RVU and mobile network data. While the number of trips according to RVU has remained relatively stable, the number of trips according to mobile network data has increased. Because the trends differ in direction in some years, the correlation between the two sources is negative, with a value of –0.20.

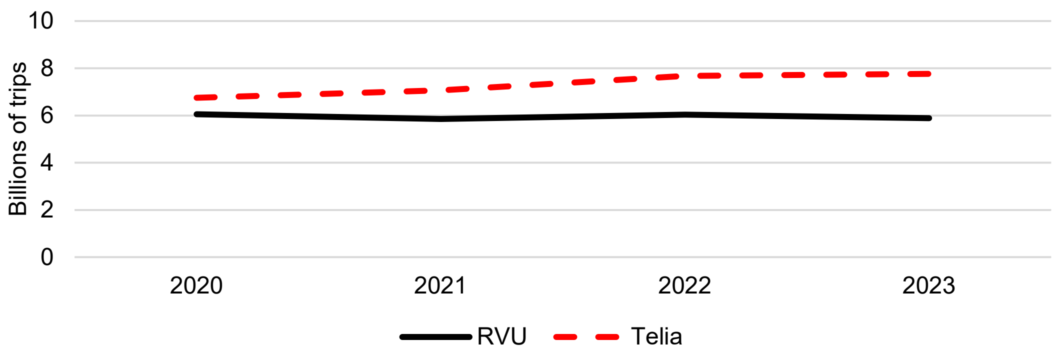

Figure 1. Number of trips according to RVU and number of trips according to Telia's mobile network data, in billions, 2020–2023.

[6] Response in an email to Trafikanalys on 9 September 2024. The original reply in Swedish has been translated into English for this memorandum.

When referring to trips in RVU, this always means trips and not journeys. A trip is a movement from an origin (stay) to the next stay, the destination. The origin and destination may have the same location or purpose, where the trip is the movement in between. The definition of trips in RVU most closely corresponds to the definition of trips in Telia's data when comparing the two sources. More information on this is provided in the previous memorandum[7].

Table 1 and the following chart in Figure 2 show the number of trips according to both sources. The analysis in the previous memorandum was based on Telia's earlier version of the data. For comparison purposes, Telia's previous data is also included here, alongside the revised figures. All analyses in this memorandum are based on Telia's updated data.

Table 1. Number of trips in billions according to the two data sources, 2019–2023.

Year	RVU (± 95% CI)	Telia's data (old)	Telia's data (new)
2019	7.0 (0.2)	10.3	7.2
2020	6.1 (0.2)	9.1	6.7
2021	5.9 (0.2)	8.9	7.1
2022	6.2 (0.3)		7.7
2023	6.0 (0.2)		7.8

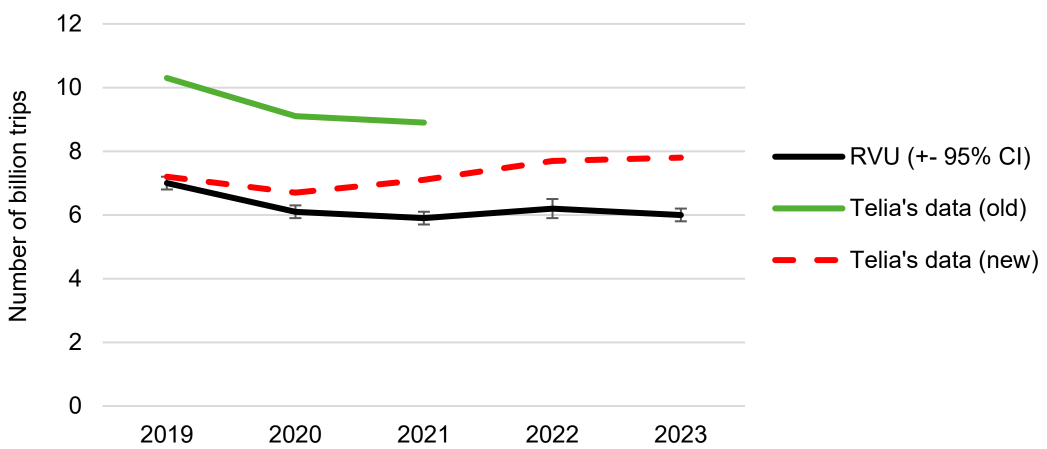

Figure 2. Number of trips according to RVU (with confidence intervals) and Telia (based on new and old methods), in billions, 2019–2023.
Note: 2019 is not comparable to other years due to a change in Telia's methodology.

As described above, Telia's data has undergone changes since the previous analysis. Despite this, we observe that mobile network data continues to show similar shortcomings to those we identified for the years 2019–2021 in our earlier study. When disaggregating the data, only the number of trips *within* counties and by month and year can be reliably estimated using Telia's data.

In addition, some travel *within* and *between* specific municipalities in metropolitan areas can be presented. However, in this project we have chosen to focus on the number of trips *within* counties and by month.

[7] Trafikanalys, "PM 2023:6 Hur väl kan mobilnätsdata beskriva våra resvanor?"

From this point forward in the project, we refer to Telia's mobile network data simply as Mobile network data, except in figures where the label "Telia's data" is retained.

3.2 Conditions

Below is a summary of the conditions for our two data sources and the intended synergy between them.

We aim to determine how many national trips are made by the population in Sweden during a given time period. The youngest individuals fall outside the population of interest, resulting in the target population being defined as individuals aged 6–84 years in RVU, and individuals aged 6 and older in the Mobile network data. The total number of trips, which we define as Y, is therefore given by:

$$Y = \sum_{\{k \in U\}} \sum_{t=1}^{T} y_{kt},$$

where *U* is the population (6–84 years in RVU, 6+ in Mobile network data), *T* is the given time period, and y_{kt} is the number of trips made by person *k* during time interval *t*.

To estimate the number of trips during a given period from RVU data, we use the estimator $\hat{Y}$, defined as:

$$\hat{Y} = \sum_{k \in U} \sum_{t=1}^{T} \delta_{kt} w_{kt} y_{kt},$$

where:

- δ_{kt} = 1 if person *k* is surveyed at time *t*, otherwise 0
- w_{kt} = weight for person *k* at time *t*
- y_{kt} = number of trips by person *k* at time *t*

To estimate the number of trips from Mobile network data, we first have the number of mobile phones observed during a given time period:

$$X_t = \sum_{d \in D} x_{dt},$$

where:

- *D* = the population of mobile phones
- x_{dt} = is mobile phone *x* at time *t*

The registrations of these phone signals are weighted by Telia to represent the population. The estimation of the total number of trips for a given period based on Mobile network data can then be described as:

$$X_t^U = \sum_{k \in U} \alpha_{kt} y_{kt} w_{kt},$$

where:

- α_{kt} = number of detected mobile phone signals
- y_{kt} = number of trips by person *k* at time *t*
- w_{kt} = Telia's weighting factor to represent the population

Ideally, the estimates from RVU and Mobile network data per year would correspond to each other and to the actual number of trips, expressed as:

$$\hat{Y} = \sum_{k \in U} \sum_{t=1}^{T} \delta_{kt} w_{kt} y_{kt} = X_t^U = \sum_{k \in U} \alpha_{kt} y_{kt} w_{kt} = Y$$

Such consistency would strengthen the case for RVU alone, and there would be no need to use Mobile network data in this analysis. Of course, it is unrealistic to expect the two sources to yield exactly the same estimates. Therefore, further analysis is needed to determine the best way to combine them.

It is important to note that RVU collects information directly from individuals about their trips and destinations, whereas Mobile network data is based on signals between mobile phones and base stations. Both sources contain uncertainties; errors may occur in how data is recorded or interpreted in both cases.

Sections 5.1 and 5.2 describe how we test and model the estimates using various methods, disaggregated by month and county. For more on the methods, see Chapter 4.

4 Method

In a previous project, we analysed the data sources to explore how Mobile network data could be used in combination with RVU. We concluded that Mobile network data may be suitable for reporting the number of trips on a monthly basis, as well as for reporting within-county trips on an annual basis. At the same time, we know that Mobile network data cannot be used on its own to replace RVU data, due to both a lack of transparency and uncertainty regarding how the Mobile network data is produced. The aim in this memorandum has therefore been to identify an appropriate method for combining the two data sources, where appropriateness refers to both quality aspects and robustness.

However, before the two data sources could be combined, further analyses were required. These analyses focused on the possibilities of reporting the number of trips disaggregated by time and geography, as well as how such an implementation should be carried out.

We also wanted to test the feasibility of estimating the number of trips by mode of transport and purpose, based on the RVU data available for those two variables.

In Section 5.1, we tested whether mobile network data correlates with RVU data when disaggregated by time (month), in order to assess which implementation method is most appropriate. In Section 5.2, we carried out a comparable analysis for disaggregation by geography (county). In Section 5.3, we compared RVU with other data sources broken down by mode of transport. The methods we used to combine and analyse the data sources are described below.

4.1 Linear prediction

One method for combining the two data sources and thereby estimating the number of trips is to use linear prediction. We have RVU data with sampling error, and we have Mobile network data.

A linear mixed model can then be expressed as:

$$\hat{Y}_t = \beta_0 + \beta_1 X_t + \upsilon_t + e_t,$$

where $\hat{Y}_t$ in our case is the estimated number of trips in RVU for a given month or a given county in a given year, β_0 and β_1 are regression coefficients, where β_0 = the intercept and β_1 = the slope coefficient, X_t = the number of trips in the Mobile network data for a given month or county in a given year, t = the month or county, υ = a structural random effect and e = the random sampling error.

To construct the best linear prediction-based (LP) estimate, we can, according to Zhang[8], use the Fay and Herriot model[9] for “small area estimation” to estimate the groups (month or county). Fay and Herriot use a variant of a James-Stein[10] estimator to reduce the variance by shrinking individual estimates towards the model estimate. In our case, we shrink the observed RVU values towards the model estimate.

[8] Zhang, ”Disaggregation of trips using MNO data”.
[9] Fay, R.E. and Herriot, R.A, ”Estimates of income for small places: An Application of James-Stein Procedures to Census Data”.
[10] James and Stein, Berkeley Symp. on Math. Statist. and Prob., ”Estimation with Quadratic Loss”.

The aim is to derive the distribution over months and counties, in the form of proportions per month and county. We can then use the formula:

$$\widehat{\hat{Y}_t}^{LP} = \hat{Y}\hat{p}_t^{LP},$$

where:

$$\hat{p}_t^{LP} = \frac{\hat{\gamma}_t \hat{Y}_t + (1-\hat{\gamma}_t)(\hat{\beta}_0 + \hat{\beta}_1 X_t)}{\sum_{l=1}^{T} \hat{\gamma}_l \hat{Y}_l + (1-\hat{\gamma}_l)(\hat{\beta}_0 + \hat{\beta}_1 X_l)},$$

where the weight γ_t depends on the degree of uncertainty in the model and in the sample:

$$\gamma_t = \frac{\sigma_v^2}{\sigma_v^2 + V(e_t)},$$

where $\sigma_v^2 = \mathrm{V}(\upsilon_t)$.

Summary:

- If the sampling error $V(e_t)$ is large → we rely more on the model estimate.
- If the model error σ_v^2 is large → we rely more on the estimate from RVU data, $\hat{Y}_t$.

4.2 Transfer learning

If there are outliers that significantly affect a linear model – so-called influencers – this indicates that model-based prediction is risky. In such cases, the method Transfer learning may be a better alternative for the estimates than linear prediction.

Transfer learning is a method within machine learning and statistics in which a model trained on one task is used as a starting point to improve performance on another, related task. Instead of training a model from scratch for every new task, Transfer learning uses the knowledge already learned from one task to improve the results.

The Transfer learning method in our project yields an estimator that resembles a James-Stein estimator[11]. This method aims to minimise the total mean squared error (MSE) across all estimators. In our case, this refers to monthly estimators and county-level estimators of the number of trips. To achieve this, the method permits a minor degree of bias for each estimator, while reducing the variance of alternative unbiased estimators. In this project, Transfer learning means that we use RVU's estimators as unbiased, but they may have relatively large variances due to the limited sample size.

Transfer learning with a James-Stein estimator helps us weight RVU and Mobile network data using a shrinkage factor applied to RVU data. We adjust the proportions (monthly and county-level) from RVU towards the proportions observed in Telia's data. The goal is to make an optimal trade-off between the weight placed on RVU estimators versus the alternatives, which are biased but have negligible variances and in this case are derived from Mobile network data, without assuming a direct linear relationship between the two sources (as is done in Linear prediction). The weight we place on RVU, and thereby also on Mobile network data, depends on how precise the estimates from RVU data are and how well the two sources correlate with each other. To illustrate the method in our project:

[11] James and Stein, Berkeley Symp. on Math. Statist. and Prob., "Estimation with Quadratic Loss".

Instead of applying equal weight to the two sources to estimate the number of trips in Stockholm County for a given year, it would be better to take into account the optimal estimates for the other counties. This is especially advantageous when aiming to reduce the overall MSE for county-level estimates.

Below is the methodology for Transfer learning from a mathematical perspective. The method[12] combines a traditional estimator, RVU's estimator $\hat{Y}_t$, with a proxy variable X_t/X, in our case Mobile network data, using a weighting parameter $\gamma \geq 0$. This parameter determines how much weight is given to each data source. Note the difference from the weight γ_t i $\hat{p}_t^{LP}$, where weights differed by month or county, whereas here the total weight is constant across all months or counties.

- If the proxy variable X_t/X exactly matches the true proportion Y_t/Y, Transfer learning provides a perfect estimate without sampling error, compared to Linear prediction, where sampling error always occurs.
- If X_t/X and Y_t/Y are completely uncorrelated, Transfer learning provides no improvement.

We now want to optimise the weighting of the two sources using Transfer learning. The method is described in three parts:

Part 1)

If we only have RVU data and want to estimate $p_t = Y_t/Y$, we want to minimise the negative log-likelihood $-\sum_{t=1}^{T} \hat{Y}_t \log p_t$; in other words, we want to maximise the likelihood of the estimate. The solution, according to Zhang, then becomes:

$$-\sum_{t=1}^{T} \hat{Y}_t \log p_t + \lambda \sum_{t=1}^{T} (p_t - 1)$$

under the constraint that $\sum p_t = 1$, where λ = the Lagrange multiplier to satisfy the constraint. The solution for the estimate is then $\hat{p}_t = \hat{Y}_t/\hat{Y}$.

Part 2)

We now want to adjust the probabilities using proxy data, i.e. Mobile network data in our case. Zhang describes how we can apply Transfer learning, where the proxy data $q_t = X_t/X$ is incorporated with a weighting parameter γ, which introduces a penalty in the optimisation function for deviating from q_t. The new optimisation problem combines:

- Likelihood from the RVU estimate $\hat{Y}_t$

and

- A divergence (Kullback–Leibler) between p_t and q_t.

This means minimising $-\sum_{t=1}^{T} \hat{Y}_t \log p_t$ through:

$$-\sum_{t=1}^{T} \hat{Y}_t \log p_t + \gamma \sum_{t=1}^{T} X_t (\log q_t - \log p_t)$$

with the constraint $\sum p_t = 1$.

With Transfer learning[13], we can then show:

[12] Zhang, "Disaggregation of trips using MNO data"
[13] Zhang, "Disaggregation of trips using MNO data"

$$\hat{p}_t = \psi(\gamma) \cdot \frac{\hat{Y}_t}{\hat{Y}} + \big(1 - \psi(\gamma)\big) \cdot \frac{X_t}{X},$$

where:

$$\psi(\gamma) = \frac{\hat{Y}/X}{\gamma + \hat{Y}/X}$$

Summary of Part 2:

- If the weighting parameter $\gamma = 0$→ we use only RVU's estimate $\hat{Y}_t$.
- If the weighting parameter $\gamma \to \infty$→ we rely entirely on the proxy data (Mobile network data) X_t/X.

Part 3)

In the third step, we describe how, according to Zhang, to choose the optimal weighting factor $\psi(\gamma)$, which controls the balance between:

- $\hat{p}_t = \hat{Y}_t/\hat{Y}$ (RVU data)

and

- $q_t = X_t/X$ (Mobile network data)

We examine the expected MSE between the estimate $\hat{p}_t$ and the true p_t, where $u_t = q_t - p_t$ measures how much the Mobile network data deviates from the true proportion. However, the true proportion p_t is not known in practice.

The formula for expected MSE is:

$$E\left(\sum_{t=1}^{T}(\hat{p}_t - p_t)^2\right) = \psi^2 \sum V(\hat{p}_t) + (1-\psi)^2 \sum u_t^2$$

Since p_t is not known in practice, ψ is replaced by an estimated version $\hat{\psi}$, based on observed data. To minimise the expected MSE, we obtain according to Zhang[14]:

$$\hat{\psi} = \frac{\hat{\tau}_u}{\hat{\tau}_u + \hat{\tau}_\varepsilon},$$

where:

- $\hat{\tau}_u = \frac{1}{T}\sum(\hat{p}_t - q_t)^2 - \hat{V}(\hat{p}_t)$ (which replaces $\frac{1}{T}\sum_{t=1}^{T} u_t^2$ since p_t is not known)
- $\hat{\tau}_\varepsilon = \frac{1}{T}\sum \hat{V}(\hat{p}_t)$

This estimate of the minimum MSE is then used to calculate an improved TL estimate $\tilde{p}_t^{TL}$.

Summary of Part 3:

- The optimal weighting between the two data sources (RVU and Mobile network data) is determined by $\hat{\tau}_u$, i.e. how well the Mobile network data aligns with the RVU data, and $\hat{\tau}_\varepsilon$, i.e. how uncertain the estimate $\hat{Y}_t$ from RVU is.

[14] Zhang, "Disaggregation of trips using MNO data"

- The better the alignment between the sources, the lower the value of $\hat{\psi}$, which according to the formula

$$\hat{p}_t = \psi(\gamma) \cdot \frac{\hat{Y}_t}{\hat{Y}} + \left(1 - \psi(\gamma)\right) \cdot \frac{X_t}{X}$$

gives more weight to the Mobile network data.

- The more reliable the estimate from RVU data $\hat{Y}_t$, the higher the value of $\hat{\psi}$, and thus more weight is placed on RVU.

4.3 Linear prediction combined with Transfer learning

To achieve an even more robust method than Transfer learning alone, we can, provided that no influencer strongly affects the model, advantageously combine linear prediction and Transfer learning in a single model. A model for applying both linear prediction and Transfer learning as a method[15] can, according to Zhang, be written as:

$$\hat{p}_t^{MX} = \omega\hat{\mu}_t + (1 - \omega)q_t,$$

where $\hat{\mu}_t = \hat{p}_t - q_t$, and the weight ω depends on how much the estimates of proportions per month or county from RVU ($\hat{p}_t$) and Mobile network data (q_t) differ from each other:

$$\omega = \frac{\sum_t (\hat{p}_t - q_t)^2}{\sum_t (\hat{p}_t - \mu_t/Y)^2 + \sum_t (\hat{p}_t - q_t)^2}$$

4.4 Summary of methods

In addition to the models where we combined the sources, we have also tested to describe the number of trips using each source separately.

In summary, we have tested the following five different methods to describe the number of trips per month and the number of trips within counties:

a) RVU data only ($\hat{Y}$)
b) Mobile network data only (X)
c) RVU data combined with Mobile network data using estimates from a linear mixed model (LP)
d) RVU data combined with Mobile network data using a Transfer learning approach (TL)
e) RVU data combined with Mobile network data using a combination of LP+TL (MX estimator)

It would be desirable to have access to data from several, preferably all, mobile operators in order to obtain a more comprehensive picture of the number of trips. This would enable cross-analysis and make it easier to determine whether changes or discrepancies are due to actual changes in travel behaviour or differences between data sources. However, we do have access to estimated trips from other products produced by Transport Analysis, such as

[15] Zhang, ”Disaggregation of trips using MNO data”

Regional scheduled public transport and *Vehicle mileage for Swedish-Registered Vehicles*[16], which we can use to analyse and understand any differences between data sources.

In Section 5.3, we have analysed the results from these products in comparison with RVU data to assess whether RVU proportions are appropriate to use for estimating the number of trips by mode of transport and purpose. The method for combining RVU data with Mobile network data – if RVU's proportions were found to be of sufficient quality – was direct weighting based on RVU proportions. In other words, a simple method was used to estimate the number of trips by mode of transport and purpose within the scope of this project.

Our analyses and subsequent methods produced distributions of the number of trips per month, per within-county, and by mode of transport and purpose on an annual basis. The number of trips was then weighted to match the total number of trips per year in the Mobile network data.

[16] Trafikanalys, " Regional scheduled public transport and Vehicle mileage for Swedish-Registered Vehicles"

5 Data analysis and testing

In this chapter, we describe how our analyses provide information on which method is most appropriate for combining the two data sources, disaggregated by month in Section 5.1 and by county in Section 5.2. Each section begins with an account of how well the data sources correlate with each other, followed by an assessment of how well the different sources or methods perform in describing the number of trips.

In Section 5.3, we describe how analyses of other statistics, together with RVU, provide insight into which method is most appropriate for combining the two data sources disaggregated by mode of transport and purpose of travel.

5.1 Correlation and testing between RVU and Mobile network data, broken down by month

Figure 3 shows the number of trips according to the two sources, covering the months of the years 2020–2023. Mobile network data is represented by a dashed red line and RVU data is shown with a solid black line. The graph shows that the two data sources follow each other relatively well when disaggregated by month, which confirms a continued strong correlation between them.

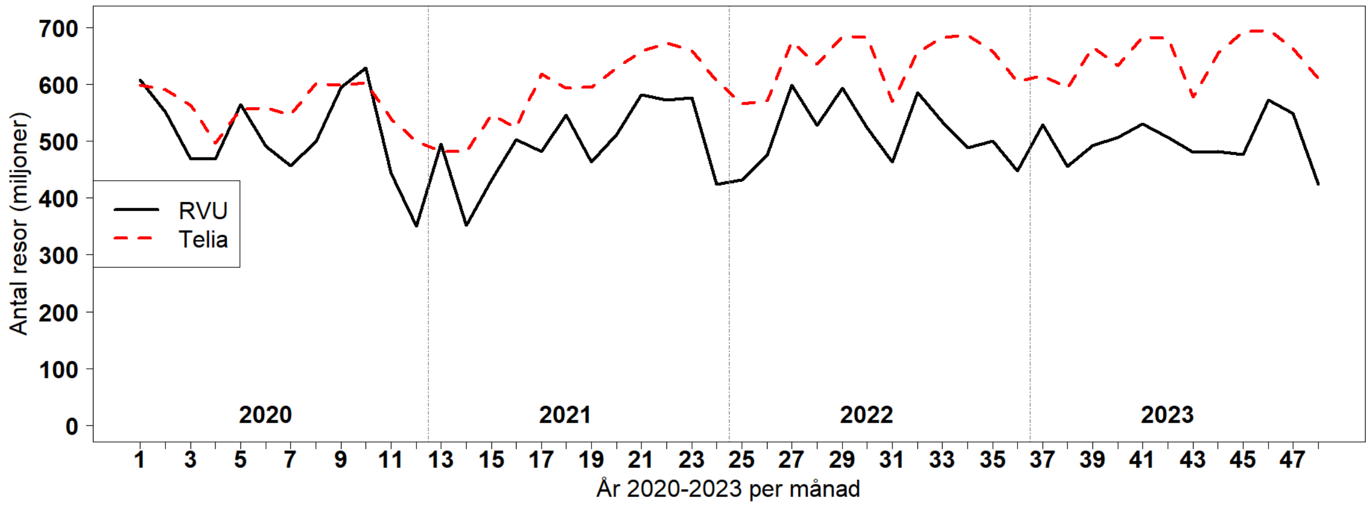

Figure 3. Number of trips in millions (y-axis) per year and month (x-axis) for the two sources, 2020–2023. 1 = January 2020, 48 = December 2023.

To determine the most suitable method for combining the two data sources by month, we tested how Mobile network data and RVU data relate to each other per year[17]. All tests in this section were carried out using the R software.

We define Y as RVU data and X as Mobile network data. The monthly RVU estimators for the number of trips are approximately normally distributed and have estimated variances. The

[17] Zhang, "Disaggregation of trips using MNO data".

RVU monthly estimators are treated as unbiased. The number of trips in the Mobile network data is treated as constant, with negligible variance but with bias. For the hypothesis testing, we used χ^2-tests (Chi-squared tests). The null hypotheses and associated testing can be expressed as follows:

a) $\mathrm{H_0}: \mathrm{X_t} = \mathrm{Y_t}$, which in words becomes $\mathrm{H_0}$: The number of trips in Mobile network data is equal to the number of trips in RVU across all 12 months in a given year, assuming that the number of RVU trips reflects the true number.

b) $\mathrm{H_0}: \mathrm{X_t} \propto \mathrm{Y_t} \Leftrightarrow \frac{\mathrm{X_t}}{\mathrm{X}} = \frac{\mathrm{Y_t}}{\mathrm{Y}}$, which in words becomes $\mathrm{H_0}$: The number of trips in Mobile network data is proportional to the number of trips in RVU across all 12 months in a given year, assuming that the distribution of trips in RVU reflects the true distribution.

In other words, we test whether it is statistically significant that Mobile network data differs from RVU data by month and year, or whether the differences fall within the sampling uncertainty of RVU.

High p-values are desired, as they indicate that the sources do not differ with respect to the number of trips and their monthly distribution. The desired outcome is that we are unable to reject the null hypotheses. The alternative hypothesis, H_1, is that there is a statistically significant difference between the data sources with regard to the number of trips and the monthly distribution.

The results show that the year 2020 is not satisfactory, as we reject the null hypotheses at all tested significance levels (α-levels). The year 2021 shows partially satisfactory results, with p-values greater than 0.1 for both null hypotheses a and b. The years 2022 and 2023 show positive results with high p-values indicating that we cannot reject the null hypotheses, which is desirable. This means that for the years 2022 and 2023, we find no statistical difference in the number of trips or their monthly distribution between the two data sources. See the p-values in Table 2.

Table 2. p-values for hypothesis test a) The number of trips in Mobile network data = the number of trips in RVU across all 12 months in a given year, and for hypothesis test b) The number of trips in Mobile network data is proportional to the number of trips in RVU across all 12 months in a given year.

Year	Null hypothesis a)	Null hypothesis b)
2020	5.08×10^{-6}	1.12×10^{-6}
2021	0.12	0.15
2022	0.60	0.63
2023	0.60	0.74

We also tested Transfer learning (TL) as a method for estimating the number of trips per month. For a description of how we applied the Transfer learning method, see Section 4.2. For this test, we analysed the measure Relative efficiency (TL) = MSE(TL) / Var(RVU). This metric indicates how efficient the TL method is compared to the direct RVU estimator, where a lower value means higher efficiency and thus a gain in precision.

Table 3 shows the relative efficiency of the TL estimator compared to RVU estimates, based on repeated sampling (100,000) with replacement, known as bootstrapping, under RVU's sampling design, where we use RVU's variances.

Table 3. Relative efficiency with a Transfer learning estimator compared with RVU estimates, disaggregated by month.

Year	Relative efficiency
2020	0.24
2021	0.077
2022	0.00039
2023	0.0099

The lower the relative efficiency in this Transfer learning method (TL), the more satisfactory the result. This method provides more satisfactory values for all four years compared with using RVU as the sole data source, including the associated uncertainty intervals. For example, a value of 0.24 for 2020 means that the variance of the TL estimates is only about one quarter of the variance of the RVU estimates. In our earlier hypothesis tests (concerning Mobile network data = RVU and Mobile network data ∝ RVU), we concluded that the results for 2020 were not satisfactory and that the 2021 p-values were not fully satisfactory. With the TL method, the results improve significantly when combining Mobile network data and RVU, when disaggregated by month and year, compared with using RVU alone.

For 2022 and 2023, we observe very low values of relative efficiency (TL), signifying greater efficiency. However, even more robust models may exist. Low relative efficiency values do not guarantee that a model is robust. To improve the robustness of the results, we further develop the model by combining a linear mixed model with Transfer learning.

The linear model is estimated with linear prediction; we refer to this method as LP in this memorandum. In addition, we test Transfer learning (TL) and, finally, a robust estimator, referred to as model MX, which combines LP and TL. By robust, we mean in this context that both linear prediction and Transfer learning are combined into one model (MX), instead of relying solely on Transfer learning.

MSE is used to evaluate the performance of an estimator. When comparing MSE between different data sources and models – as in our case, where we have RVU, Mobile network data, and combinations of the two – it is more useful to use Average Relative Root Mean Square Error (ARRMSE). This normalises the actual values measured by MSE, meaning that the scale of the number of trips does not affect the result, and a relative comparison between sources is made. We calculate ARRMSE for each method in the following order: RVU (as a single source), Mobile network data (as a single source), LP (estimates from a linear mixed model), TL (Transfer learning), and MX (robust estimator combining LP and TL). The results are presented in Table 4. Estimates of ARRMSE were produced through bootstrapping (100,000) under RVU's sampling design, for all five methods.

Table 4. Average Relative Root Mean Square Error (ARRMSE) for trips disaggregated by month, per source/method and year.

Year	RVU	Mobile network data	LP	TL	MX
2020	0.052	0.073	0.026	0.022	0.024
2021	0.083	0.026	0.022	0.021	0.022
2022	0.082	0.0017	0.000059	0.0014	0.00066
2023	0.085	0.0093	0.000079	0.0077	0.0033

As mentioned, ARRMSE is a measure of estimator performance, where a lower value indicates better precision. We observe that RVU has the highest ARRMSE value in 2021–2023 compared to Mobile network data and the other methods. This means that direct use of RVU gives the least satisfactory results in terms of the highest mean squared error when disaggregated by month for the years 2021–2023. We also note that RVU has the lowest ARRMSE in 2020, even slightly lower than Mobile network data that year, compared with the other years. Mobile network data and the other methods show clearly lower ARRMSE in 2022 and 2023 compared with the first two years.

Figure 4 – Figure 7 show the results by year. The black data points represent a linear model where both data sources are integrated, the red dashed line shows Mobile network data, and the solid black line shows RVU, for January through December each year. We identify some outliers, but none of these are considered influencers (outliers that have a significant effect on a linear model).

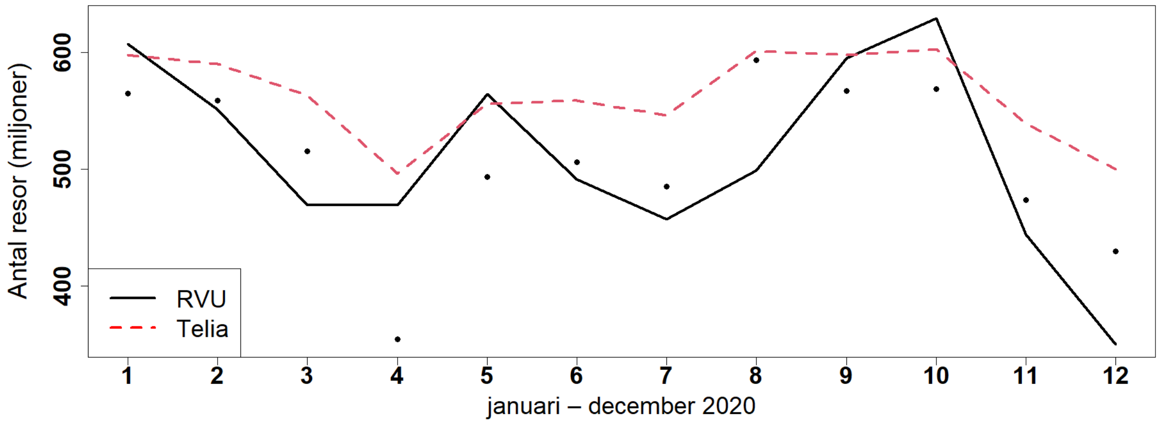

Figure 4. Number of trips in millions (y-axis) per month in 2020 (x-axis), where 1 = January, 12 = December. The black data points in the diagram indicate the estimates for a linear model.

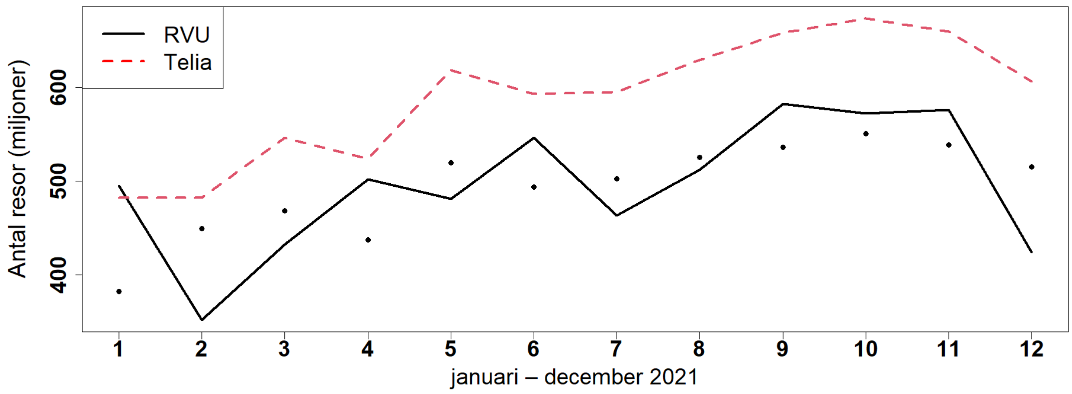

Figure 5. Number of trips in millions (y-axis) per month in 2021 (x-axis), where 1 = January, 12 = December. The black data points in the diagram indicate the estimates for a linear model.

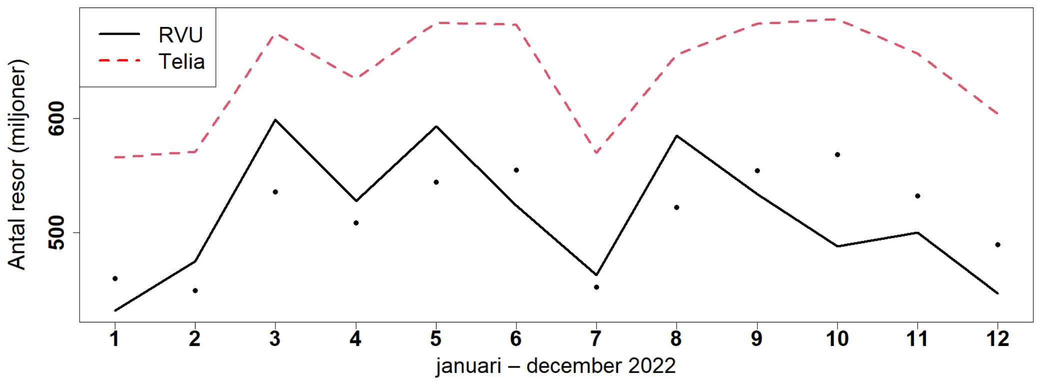

Figure 6. Number of trips in millions (y-axis) per month in 2022 (x-axis), where 1 = January, 12 = December. The black data points in the diagram indicate the estimates for a linear model.

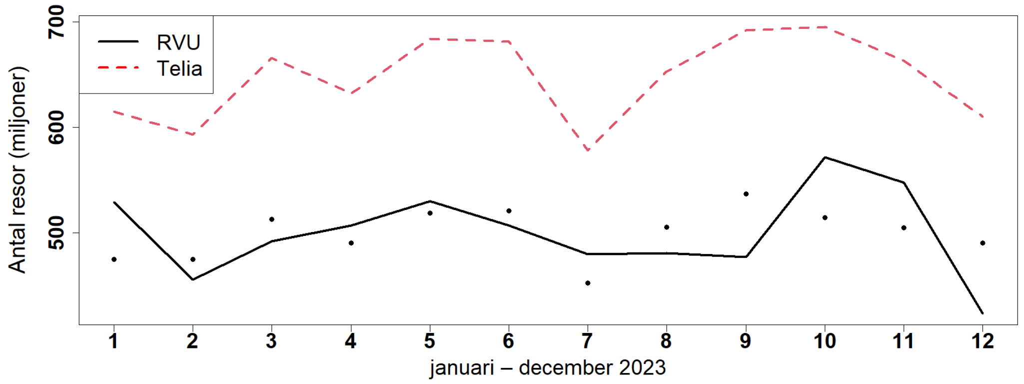

Figure 7. Number of trips in millions (y-axis) per month in 2023 (x-axis), where 1 = January, 12 = December. The black data points in the diagram indicate the estimates for a linear model.

To support the inclusion of the linear model in the method, the values of the intercept or slope coefficient of the regression line should not show significant variation between months. This means that we do not want to see large deviations for any single month.

Table 5 presents how the intercept (β_0) and slope coefficient (β_i) change when individual observations are excluded, for the years 2020–2023.

Table 5. The values for the rows β_0 show how much the intercept changes, and the values for β_i show how much the slope changes if the month is excluded, in a linear model for trips disaggregated by month. Years 2020–2023. The columns represent the months in order, where 1 = January and 12 = December.

Year, parameter	1	2	3	4	5	6	7	8	9	10	11	12
2020, β_0	-0.10	0.02	0.00	0.62	0.05	0.00	-0.03	0.27	-0.06	-0.17	-0.05	-0.40
2020, β_i	-0.05	0.01	0.00	0.28	0.02	0.00	-0.01	0.14	-0.03	-0.08	-0.02	-0.18
2021, β_0	-2.14	1.85	0.30	-0.77	-0.14	-0.01	0.00	-0.06	0.48	0.28	0.40	-0.15
2021, β_i	0.35	-0.29	-0.03	0.12	0.03	0.00	0.01	0.02	-0.08	-0.04	-0.07	0.05
2022, β_0	-1.79	1.55	-1.71	0.12	-1.69	1.04	0.65	-0.70	0.70	3.01	0.39	-1.38
2022, β_i	-0.08	0.08	-0.09	0.00	-0.09	0.06	0.03	-0.04	0.04	0.17	0.03	-0.06
2023, β_0	-0.45	0.27	-0.08	-0.06	0.10	-0.10	-0.48	-0.01	-0.64	0.67	0.15	0.66
2023, β_i	0.17	-0.10	0.04	0.02	-0.03	0.05	0.19	0.02	0.28	-0.27	-0.06	-0.24

For example, the table shows that if April 2023 (column 4) is excluded, the intercept decreases by 6 percent and the slope coefficient increases by 2 percent. We observe no clear pattern in these deviations in Table 5, which applies to all four years.

Brief summary:

- The results from bootstrapping of relative efficiency indicate that Transfer Learning (TL) is more efficient than RVU in terms of variance.
- When comparing the five methods using bootstrapping (RVU only, Mobile network data only, LP, TL, MX), we see that MX consistently yields the lowest ARRMSE value. The MX model thus indicates better precision than RVU on an annual level when compared with RVU's uncertainty estimates.
- Based on visual analysis of Figure 4, Figure 5, Figure 6 and Figure 7, as well as Table 5, we conclude that the linear model can be included in our modelling.
- If estimates in a linear mixed model can be combined with Transfer learning (LP+TL=MX) without linear prediction becoming uncertain, this provides an advantage in the form of a more stable estimate compared with using Transfer learning alone.

We therefore consider the most robust model for estimating the number of trips per month to be the MX estimator, which combines estimates from the linear mixed model (LP) and Transfer learning (TL).

Figure 8 presents the number of trips per month for each method and data source for the year 2023. MX is the method we have chosen for estimating monthly trips, and note that the number of trips has been weighted to the total number of trips in the Mobile network data. The choice of source for the total number of trips is discussed in Chapter 6.

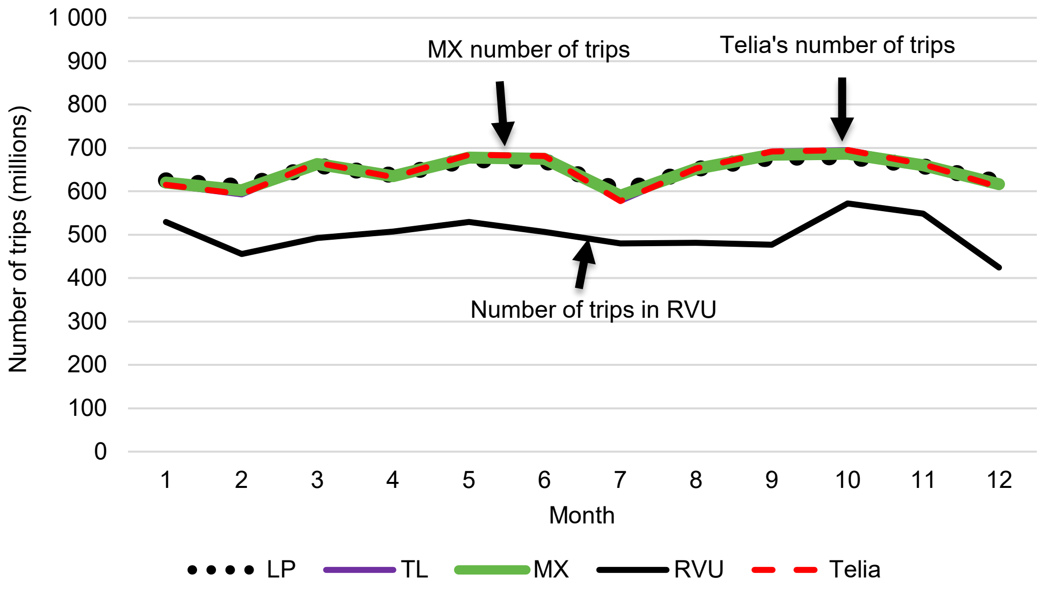

Figure 8. Number of trips per month for the two sources and for the three methods, weighted to Telia's total number of trips. Year 2023.

In Figure 8, we observe that Mobile network data follows MX relatively well during 2023. However, the results for 2020 (see Figure 9) show that RVU instead follows MX relatively well.

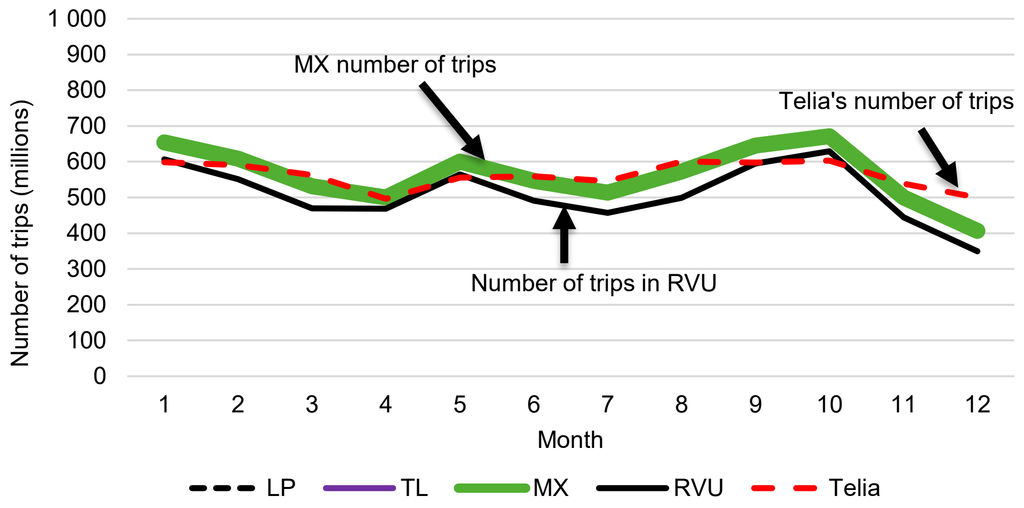

Figure 9. Number of trips per month for the two sources and for the three methods, weighted to Telia's total number of trips. Year 2020.

5.2 Correlation and testing between RVU and Mobile network data, broken down by county

Figure 10 presents a chart covering the four years 2020–2023, where the number of trips in Mobile network data is represented by a red dashed line and the number of trips in RVU by a black line. The graph displays the number of trips within counties, with the counties sorted alphabetically and the years in ascending order from left to right.

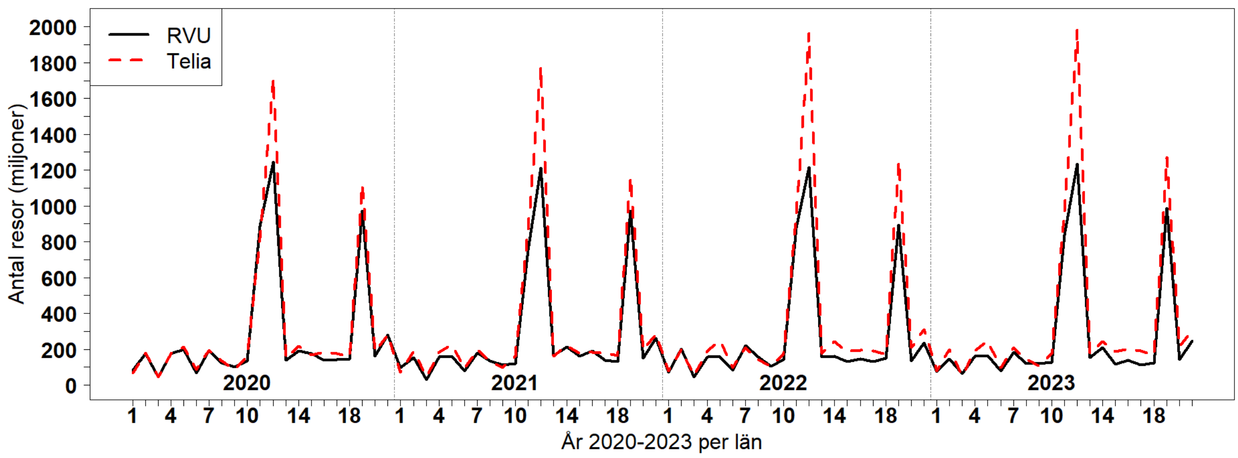

Figure 10. Number of trips in millions (y-axis) per year and within counties (x-axis) for the two sources, 2020–2023. An explanation of the county numbers on the x-axis is available in Table 11 in the Appendix.

We test how well the two data sources align with each other per year to determine the most appropriate implementation method[18]. All tests in this section were conducted using R.

[18] Zhang, ”Disaggregation of trips using MNO data”.

We define Y as RVU data and X as Mobile network data. The RVU county estimators for the number of trips are approximately normally distributed and have estimated variances. The RVU county estimators are treated as unbiased. The number of trips in the Mobile network data is treated as constant, with negligible variance but with bias. For the hypothesis testing, we used χ^2-tests (Chi-squared tests). The null hypotheses and associated hypothesis testing can be expressed as follows:

a) $\mathrm{H_0}: \mathrm{X_t} = \mathrm{Y_t}$, which in words becomes $\mathrm{H_0}$: The number of trips in Mobile network data is equal to the number of trips in RVU across all 21 counties in a given year, where only trips <u>within</u> counties are measured and we assume that RVU's number of trips reflects the true number.

b) $\mathrm{H_0}: \mathrm{X_t} \propto \mathrm{Y_t} \Leftrightarrow \frac{\mathrm{X_t}}{\mathrm{X}} = \frac{\mathrm{Y_t}}{\mathrm{Y}}$, which in words becomes $\mathrm{H_0}$: The number of trips in Mobile network data is proportional to the number of trips in RVU across all 21 counties in a given year, where only trips <u>within</u> counties are measured and we assume that RVU's distribution of trips reflects the true distribution.

The alternative hypothesis, H_1, is that there is a statistically significant difference between the data sources with respect to the number of trips and the distribution of trips per county.

In summary, we test whether Mobile network data differs from RVU with regard to trips within counties and years, or if the differences fall within the sampling uncertainty of RVU. As in Section 5.1, we want the p-values to be high because it means we cannot reject the null hypotheses. In other words, we do not want the sources to differ in terms of the number of trips and the distribution of trips within counties. However, the results are not satisfactory, as all the null hypotheses are rejected (see p-values in Table 6).

Table 6. p-values for hypothesis test a) The number of trips in Mobile network data = the number of trips in RVU across all 21 counties in a given year, and for hypothesis test b) The number of trips in Mobile network data is proportional to the number of trips in RVU across all 21 counties in a given year.

Year	Null hypothesis a)	Null hypothesis b)
2020	0.00	0.00
2021	0.00	5.61×10^{-5}
2022	0.00	2.00×10^{-7}
2023	0.00	0.00022

We also test Transfer learning (TL) to estimate the number of trips. The measure Relative efficiency (TL) = MSE(TL) / Var(RVU) is used for this. As previously mentioned, a lower value indicates greater efficiency compared with the direct RVU estimator, based on bootstrapping (100,000) under RVU's sampling design.

Table 7 shows the relative efficiency of a TL estimator compared with RVU estimates, based on bootstrapping (100,000) under RVU's sampling design.

Table 7. Relative efficiency with a Transfer learning estimator compared with RVU estimates, disaggregated by county.

Year	Relative efficiency
2020	0.029
2021	0.17
2022	0.089
2023	0.20

This method (TL) provides more satisfactory results for all four years compared with an estimate based solely on RVU as a source, including the associated uncertainty intervals. For example, a value of 0.029 for 2020 means that the MSE is only about 3 percent of the variance of the RVU estimates. This means that TL provides better precision when estimating the number of trips per county compared with using only RVU. Therefore, the TL method yields more reliable results when disaggregated by county and year than a model based solely on RVU.

As in the tests performed in Section 5.1, we want to further strengthen the robustness of the model by combining estimates from a linear mixed model with a Transfer learning method. We therefore test Linear prediction (LP), Transfer learning (TL), and finally a robust estimator, which we call model MX, which combines LP and TL.

Using Mobile network data or RVU directly gives the worst results when disaggregated by county, where Mobile network data shows slightly higher (and therefore worse) values than RVU in terms of Average Root Mean Square Error (ARRMSE). As described in Section 5.1, ARRMSE is used to evaluate an estimator's performance. Estimates of ARRMSE were produced through bootstrapping (100,000) under RVU's sampling design for all five methods.

Table 8 below presents ARRMSE for the following methods: RVU (as a single source), Mobile network data (as a single source), LP (estimates from a linear mixed model), TL (Transfer learning), and MX (robust estimator combining LP and TL). Estimates of ARRMSE were produced through bootstrapping (100,000) under RVU's sampling design for all five methods.

Table 8. Average Relative Root Mean Square Error (ARRMSE) for trips disaggregated by county, per source/method and year.

Year	RVU	Mobile network data	LP	TL	MX
2020	0.075	0.11	0.016	0.0071	0.0096
2021	0.11	0.11	0.059	0.034	0.051
2022	0.10	0.13	0.032	0.024	0.027
2023	0.11	0.11	0.044	0.040	0.042

The results show that Transfer learning (TL) and the robust method MX (which combines LP and TL) exhibit the lowest (best) ARRMSE values when estimating the number of trips within counties. This indicates that TL and MX are more advantageous than using RVU or Mobile network data alone.

Figure 11 – Figure 14 below show how a linear model performs per *within*-county in the diagrams. The red dashed line represents the number of trips in Mobile network data, the black line represents the number of trips in RVU, and the black data points correspond to a linear model where both sources are integrated.

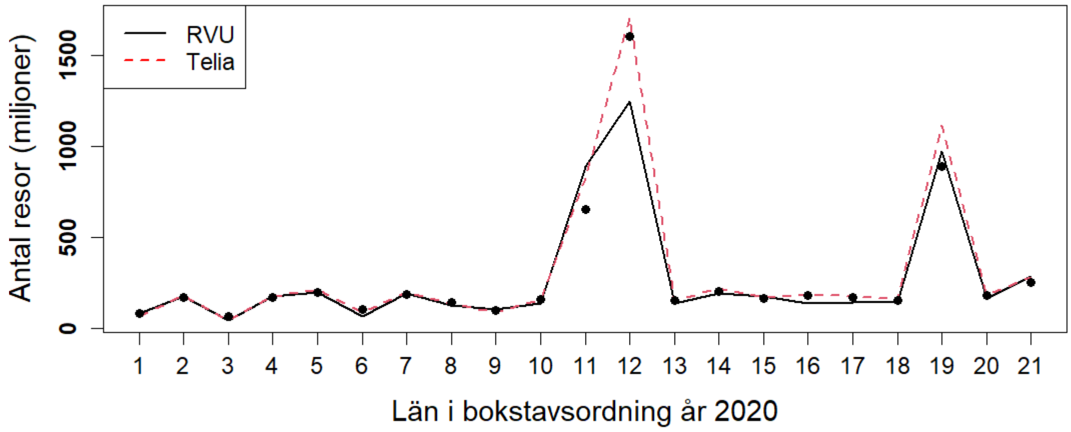

Figure 11. Number of trips in million (y-axis) per county in 2020 (x-axis). An explanation of the county numbers on the x-axis is available in Table 11 in the Appendix. The black data points in the diagram represent the estimates for a linear model.

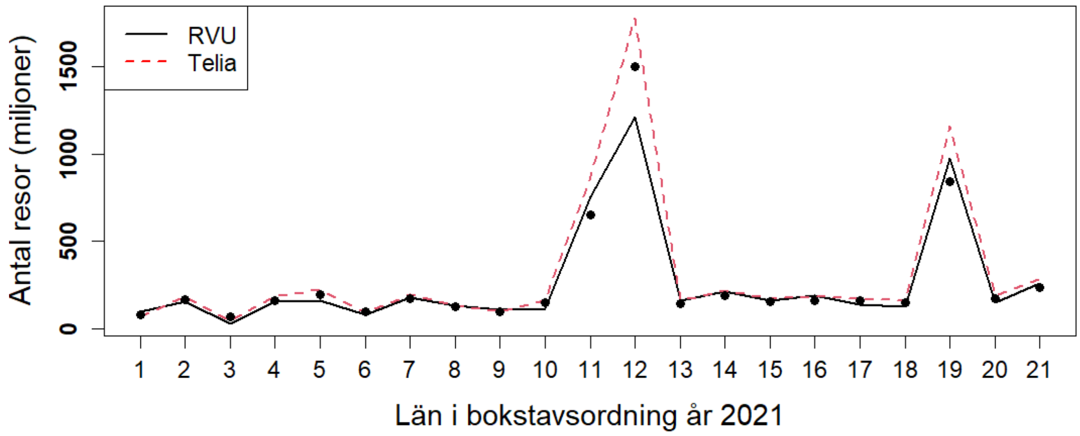

Figure 12. Number of trips in million (y-axis) per county in 2021 (x-axis). An explanation of the county numbers on the x-axis is available in Table 11 in the Appendix. The black data points in the diagram represent the estimates for a linear model.

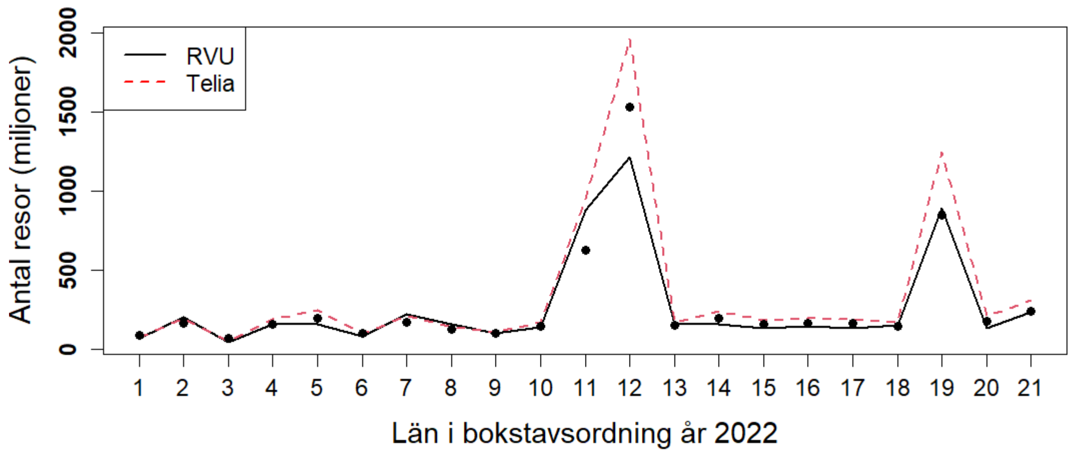

Figure 13. Number of trips in million (y-axis) per county in 2022 (x-axis). An explanation of the county numbers on the x-axis is available in Table 11 in the Appendix. The black data points in the diagram represent the estimates for a linear model.

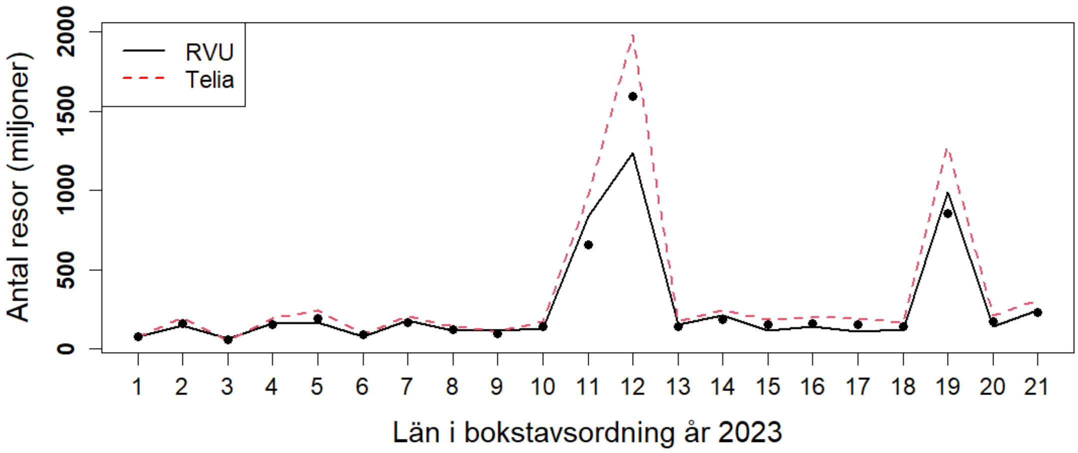

Figure 14. Number of trips in million (y-axis) per county in 2023 (x-axis). An explanation of the county numbers on the x-axis is available in Table 11 in the Appendix. The black data points in the diagram represent the estimates for a linear model.

When analysing the three counties with the most trips: Skåne County (11), Stockholm County (12), and Västra Götaland County (19), we observe that the linear model is risky. These three counties contain so-called influencers, i.e., outliers that have a significant impact on a linear model, and this applies consistently across all four years. Table 9 presents how the intercept (rows β_0) and slope coefficient (rows β_i) in the linear model change when an individual data point (county) is excluded, per year.

Table 9. The values for the rows β_0 show how much the intercept changes, and the values for β_i show how much the slope changes if the county is excluded, in a linear model for trips disaggregated by county. Years 2020–2023. The columns represent the counties in alphabetical order, where 1 = Blekinge County and 21 = Östergötland County.

Year, Parameter	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
2020, β_0	0.00	0.00	0.05	0.00	0.00	0.08	-0.01	0.03	0.00	0.05	0.00	**-0.94**	0.03	0.02	-0.01	0.08	0.05	0.02	0.07	0.04	-0.04
2020, β_i	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**-0.04**	**0.19**	0.00	0.00	0.00	0.00	0.00	0.00	**-0.02**	0.00	0.00
2021, β_0	-0.02	0.03	0.08	0.01	0.06	0.04	0.00	0.00	-0.01	0.06	0.00	**-0.71**	-0.02	-0.03	0.00	-0.04	0.05	0.04	0.11	0.04	-0.03
2021, β_i	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**-0.01**	**0.16**	0.00	0.00	0.00	0.00	0.00	0.00	**-0.03**	0.00	0.00
2022, β_0	0.03	-0.06	0.05	0.00	0.07	0.04	-0.07	-0.04	0.00	0.01	0.00	**-0.73**	0.00	0.06	0.05	0.03	0.05	0.00	0.03	0.07	0.00
2022, β_i	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**-0.04**	**0.18**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2023, β_0	0.01	0.03	0.00	0.00	0.06	0.04	-0.03	0.00	-0.05	0.04	0.01	**-1.14**	-0.01	-0.04	0.09	0.05	0.11	0.04	0.13	0.07	-0.02
2023, β_i	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**-0.03**	**0.19**	0.00	0.00	0.00	0.00	0.00	0.00	**-0.03**	0.00	0.00

For a linear model to function satisfactorily, the differences between the counties (columns) should not be too large. In this case, however, we see that column 12 (Stockholm County) in 2023 has an intercept value of –1.14 (i.e., the intercept decreases by 114 percent if this data point is excluded) and a slope coefficient of 0.19 (i.e., the slope increases by 19 percent when this data point is excluded). Furthermore, there are only three data points where the slope differs from 0. The same pattern observed in 2023 is also seen in the other three years. This clear difference between the counties presents a risk because it can lead to large errors in the model if a linear model is used in our modelling. The Transfer learning method allows for bias but is less sensitive than a linear model when the model is poorly fitted.

Transfer learning is more resistant to influencers. This is because the Transfer learning method does not require the assumption of a linear model.

Brief summary:

- The results from bootstrapping of relative efficiency show that TL is more efficient in terms of variance compared with RVU.
- We have observed that TL and MX yield the lowest ARRMSE overall when comparing all five methods (RVU only, Mobile network data only, LP, TL, MX) from bootstrapping.
- The MX method is more sensitive, and therefore more risky, than TL because it is relatively more affected when data points for the three counties with the highest number of trips are excluded.
- Based on the visual analysis of Figure 11, Figure 12, Figure 13 and Figure 14, as well as the analysis of Table 9, we conclude that the linear model cannot be included in our modelling.
- At the beginning of the section, we observed that the two data sources differ in terms of the number of trips and the distribution of trips within counties. This indicates that linear modelling may be inappropriate.

Given these results, we conclude that Transfer learning (TL) is the most suitable implementation method for estimating the number of trips disaggregated by county.

In Figure 15, the number of trips per county for each method and source for the year 2023 is presented. TL has been chosen as the method for trips per within-county. Note that the trips in the methods have been weighted to the total number of trips in the Mobile network data, similar to what was done in Section 5.1. A discussion regarding the choice of source for the total number of trips is provided in Chapter 6.

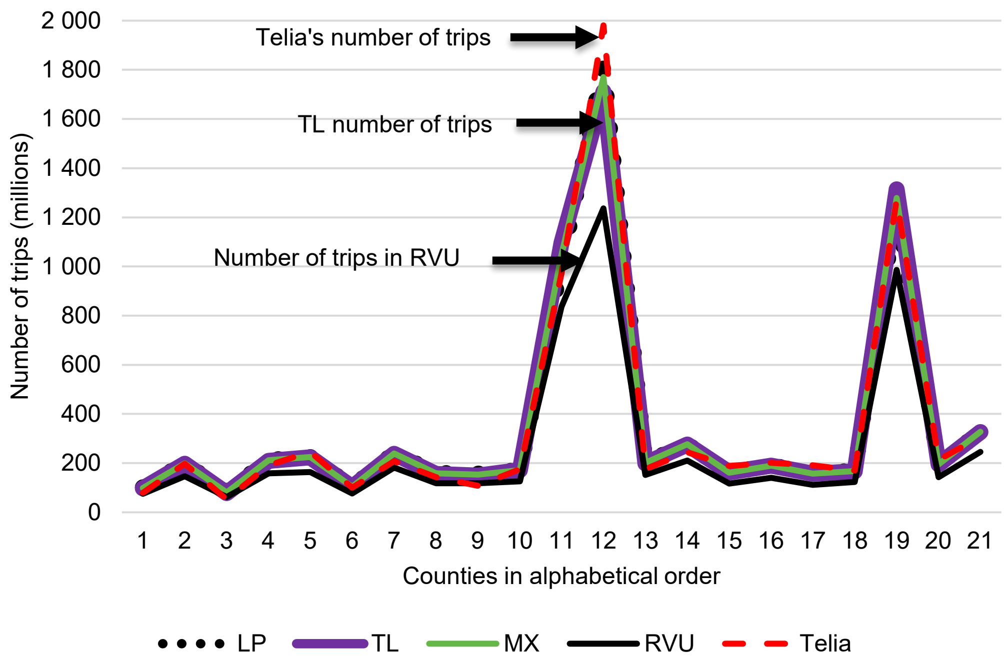

Figure 15. Number of trips per county for the two sources and for the three methods, weighted to Telia's total number of trips. Year 2023.

5.3 Correlation between RVU and other statistics

In Section 3.1 and Figure 1, we observed that the correlation between RVU and Mobile network data was very low regarding the annual development of the number of trips. However, it is difficult to determine which of the curves represents the most reasonable development. With Mobile network data, it is not possible to disaggregate travel by mode of transport. For RVU, it is possible to disaggregate trips by mode of transport, and there are other sources for comparison. For the modes of transport such as buses, subways, and trams, there is official boarding statistics. Since a trip in RVU can include several boardings, it is expected that the levels will differ. When comparing the development of boardings from 2019 onwards with RVU's data on trips, we see an expected difference in the number of trips, but the curves correlate well (the correlation is 0.99), see Figure 16.

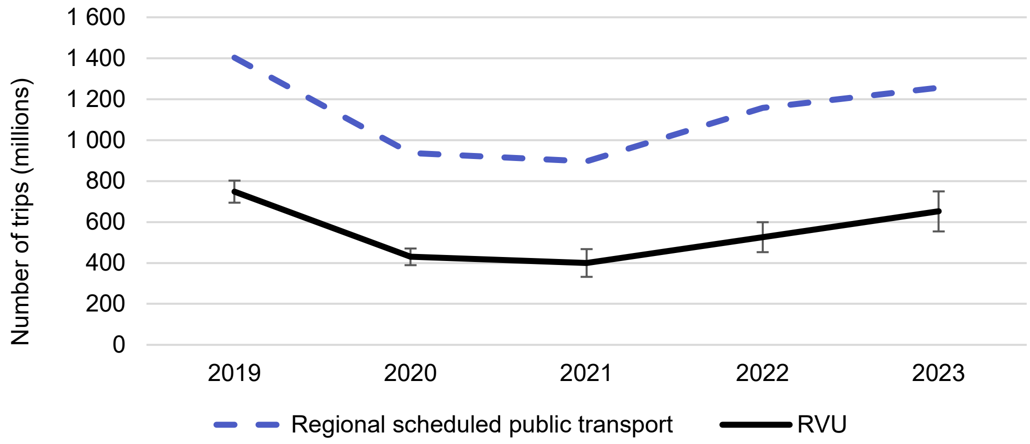

Figure 16. Number of trips in RVU and boardings according to Regional scheduled public transport in millions, 2019–2023.

Similarly, the mileage according to official statistics, which is based on car inspection data from the vehicle fleet, follows the development of mileage with passenger cars according to RVU (the correlation is 0.73), see Figure 17.

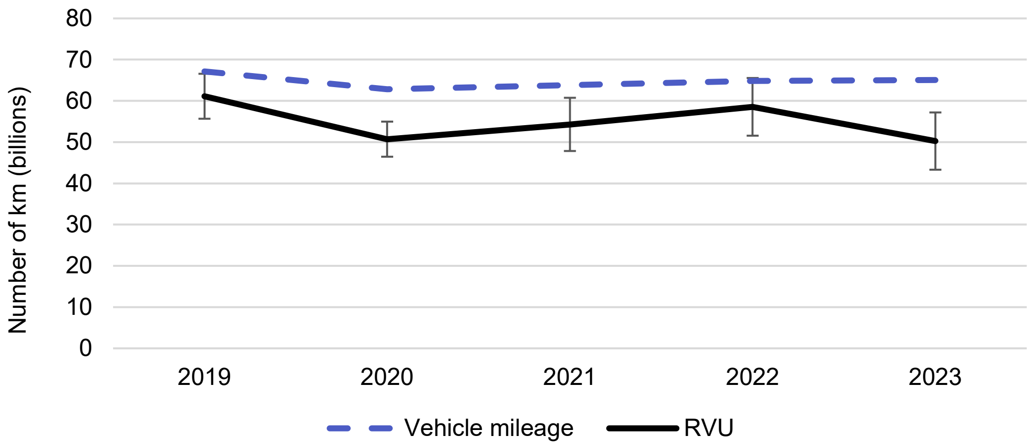

Figure 17. Distance travelled as a driver of a passenger car according to RVU and vehicle mileage for Swedish-registered vehicles according to official statistics, 2019–2023. Billions of km.

The reason that travel according to RVU has not increased much after the pandemic is mainly due to a reduction in cycling and walking. However, for these modes of transport, we do not have another source for comparison to see if the development seems reasonable. On the other hand, we observe a reasonable development for the other major modes of transport, passenger cars and public transport.

Assuming that RVU shows correct proportions for different modes of transport, it is possible to create an index of how travel has developed according to other official statistics (hereafter referred to as the index for official statistics). To create the index for official statistics, it is assumed that the number of passenger-kilometres for railway transport, boardings in regional public transport, mileage for passenger cars, and domestic air passengers in 2019 corresponds to the proportion of trips by train (railway), regional public transport, passenger car, and air according to RVU for the same year (see *Proportion excluding other* in Table 10). Thus, mileage for passenger cars has a large impact on the index for official statistics (79.4 percent), while domestic air travel has a minor impact (0.3 percent).

Since we lack official statistics for walking and cycling, the index for official statistics does not account for changes in travel with these modes, despite them representing a relatively large portion of travel (30.3 percent). This means that there is a relatively large uncertainty in the index for official statistics.

Table 10. Mode of transport proportions according to RVU 2019.

Mode of transport	Proportion	Proportion excluding other
Railway	3.4%	4.8%
Regional public transport	10.8%	15.4%
Passenger car	55.3%	79.4%
Air	0.2%	0.3%
Other	30.3%	

When the index for official statistics is compared with the development of the number of trips according to Mobile network data, the correlation is relatively high (0.98). We also find a high correlation between the index for official statistics and RVU, excluding other modes of transport than train, regional public transport, passenger car, and air (0.90).

However, RVU correlates poorly if all modes of transport are included. The correlation with Mobile network data is then –0.21 and with the index for official statistics –0.09, see Figure 18.

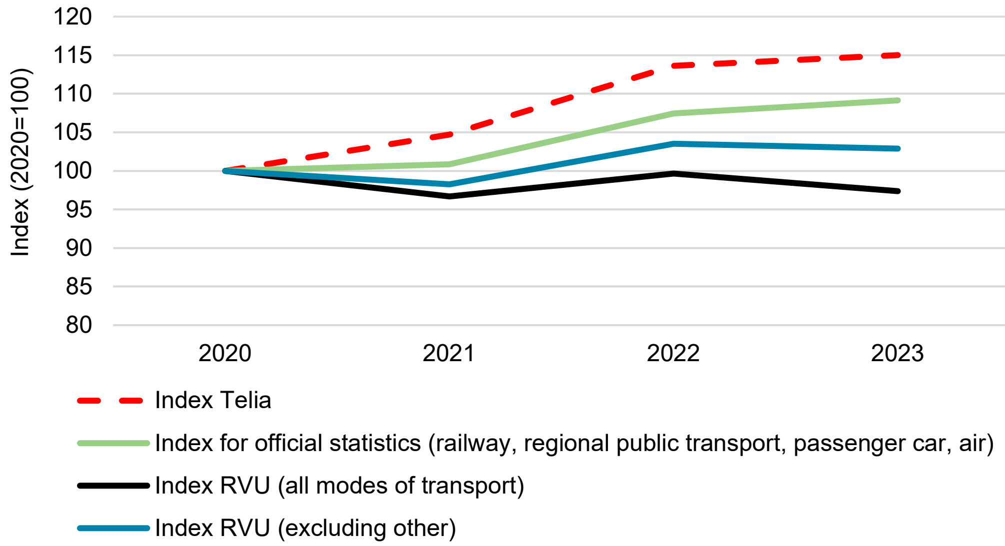

Figure 18. Development of travel according to index (2000=100) for the period 2020–2023.

The fact that RVU aligns relatively well with other comparable statistics regarding modes of transport suggests that RVU is a good source for tracking the development of travel.

In Mobile network data, as we previously noted, there is no access to either mode of transport or purpose of travel. Based on the assumption that RVU accurately captures flows by mode of transport, we have chosen to use RVU's proportions to describe travel volumes by mode of transport, annually.

The purpose of travel is also published in RVU, and this categorisation adheres to the standard for official statistics. Therefore, we have chosen to use RVU's proportions to describe travel volumes by purpose and year. Mode of transport and purpose are thus assigned RVU's proportions, and we have, as we did for breakdowns by month and county, weighted these proportions to the total number of trips in the Mobile network data per year. This method is called a direct weighting.

6 Discussion and conclusions

In Chapter 5, we analysed the most suitable implementation methods for breaking down the material by month and by county. The results from the analyses showed that methods that combine the two data sources produced more stable results compared to using the data sources separately.

For disaggregation by month, the analyses showed that a robust mixed model, which combines Transfer learning with estimates from a linear mixed model, provided the best results for estimating the number of trips using both data sources. See Section 5.1 for the analyses.

For disaggregation by county, however, it turned out that the Transfer learning method, when used individually and not together with a linear mixed model, was the most suitable for estimating the number of trips using both data sources. See Section 5.2 for the analyses. The data disaggregated by county, unlike the data disaggregated by month, contains influencers that would significantly affect a linear mixed model. Therefore, we consider it a safer method to use Transfer learning alone for disaggregation by county – safer in the sense of reducing the risk of large estimation errors in the model, which could occur with a linear model in that case.

In addition to disaggregation by time and geography, the project also included breakdowns of the number of trips by mode of transport and purpose, by combining the two data sources. We concluded that it is possible to present mode of transport and purpose using RVU, which we can then combine with Mobile network data. Through a direct weighting, we combine RVU's proportions with the total number of trips from Mobile network data to describe trips by mode of transport and purpose. See Section 5.3 for the analyses.

We concluded that the total number of trips per year is described by Mobile network data in the combined data model. One argument for using Mobile network data as a source for weighting the trips is its broader target population, which includes the population from 6 years and older, compared to 6–84 years in RVU. However, we are aware that there are uncertainties surrounding the measurement in Mobile network data, and we generally find this source to be less reliable than our product RVU. It is also a "black box" for us how the weighting up to the total number of trips in Mobile network data has been performed. We regard our presentation as the result of this project as experimental statistics and not official statistics. Therefore, we have chosen to weight to the total number of trips based on Mobile network data and not to the total number of trips based on RVU data. We also want to be clear to statistical users about the differences between the publication from this project and the publication represented by official statistics in the national travel survey, RVU. This is also a reason why we do not use RVU data to weight the total number of trips.

If we gain access to additional Mobile network data sources from more operators in the future, it may be even more advantageous to use Mobile network data as a source for weighting the total number of trips. Another alternative for weighting the total number of trips could be to combine RVU and Mobile network data.

There are disadvantages to traditional sample surveys, such as declining response rates, which contribute to greater measures of uncertainty. This becomes particularly challenging when disaggregating the statistics by time periods and geography. An advantage of Mobile network data, assuming it is accurate, is that its large data set allows for detailed disaggregation, which surpasses RVU in that regard. We are aware that there are uncertainties surrounding Mobile network data, as previously mentioned, which is the reason for the current decision not to publish statistics between geographical locations.

Based on the conclusions, we will present the following new tables:

a) Number of trips per month

b) Number of trips per within-county

c) Number of trips by mode of transport and purpose

The following estimation methods were applied for each table:

a) a robust mixed model (an MX estimator) that combines estimates in a linear mixed model (LP) with Transfer learning (TL) for disaggregation by month.

b) a method with Transfer learning (TL) for disaggregation by county.

c) a direct weighting method where we use RVU's proportions for disaggregation by mode of transport and purpose.

7 Publication of the results

The results from the work combining RVU with Mobile network data are presented through the visualisation platform used by Transport Analysis to publish statistics in the form of tables and figures, and are available at the following webpage: www.trafa.se/transportmonster/RVU-Sverige/kombinerade-mobilnatsdata-och-enkatdata-beskriver-resmonster-15129/

We present the results in the form of tables and figures for the years 2020–2023 with the following breakdowns:

a) Number of trips per month.

b) Number of trips per county.

c) Number of trips by mode of transport and purpose.

Figure 19 shows a screenshot from the publication page, only in Swedish.

Ungefär 95 procent av alla resor har sin start- och slutpunkt inom ett och samma län. Flest resor sker inom Stockholms län och minst antal resor på Gotland. Antalet resor har ökat för varje år under perioden 2020 till 2023. År 2023 hade således flest antal resor och året då pandemin bröt ut, 2020, hade minst antal resor.

Tabell 1. Antal tusen resor inom län per län. År 2020, 2021, 2022 och 2023.

Län	2020	2021	2022	2023
Stockholms län	1 417 863	1 522 583	1 639 991	1 711 320
Uppsala län	216 140	250 596	215 250	275 240
Södermanlands län	154 059	193 263	202 928	199 119
Östergötlands län	315 423	312 309	311 685	326 308
Jönköpings län	216 037	216 167	281 781	237 960
Kronobergs län	112 231	129 211	133 287	149 612
Kalmar län	139 404	158 820	199 384	156 595
Gotlands län	50 253	39 190	58 400	78 070
Blekinge län	90 915	110 871	90 524	96 935
Skåne län	989 284	910 291	1 132 181	1 096 774
Hallands län	222 323	201 296	209 623	224 244
Västra Götalands län	1 097 589	1 176 475	1 184 126	1 313 666
Värmlands län	195 468	192 964	175 013	162 074
Örebro län	181 608	182 771	180 011	195 446
Västmanlands län	163 918	159 555	195 688	165 752
Dalarnas län	197 066	187 960	258 116	197 060
Gävleborgs län	195 787	190 129	205 870	209 445
Västernorrlands län	161 677	167 974	175 456	156 997
Jämtlands län	74 869	96 156	105 718	102 519
Västerbottens län	155 812	223 681	193 667	189 919
Norrbottens län	151 286	146 302	185 049	171 054
Riket	6 499 013	6 768 564	7 333 747	7 416 109

Figure 19. Screenshot from 2025 of the visualisation platform used by Transport Analysis to publish the statistics, where the publication has been exclusively presented. The table in the image shows the number of trips per county and year using a Transfer learning method (TL), for the years 2020–2023.

8 Future work

In the future, we want to continue to further develop the statistics. This may include collecting data from more mobile network operators than the single operator, Telia, which was used in this project. By including more operators, we would be able to cross-validate mobile network data between different operators' data and RVU. This would not only provide a more extensive data set but also increase the robustness of the analyses and provide better conditions for creating more detailed and higher-quality statistics.

One area of development is to more sophisticatedly utilise the possibilities of mobile network data to break down the material into more detailed divisions and to explore relationships between different locations. To ensure that this breakdown is feasible, further analyses of the quality of mobile network data are required.

Another area for development is to better integrate variables from RVU that are not present in mobile network data, such as mode of transport and purpose. This would involve both improving and refining the current method we have used to classify trips by mode of transport and purpose, as well as adding more dimensions, for example, the breakdown of trips by gender and age.

It may become relevant to change the source for the total number of trips per year, either by using data from more mobile network operators or by combining mobile network data with RVU.

To further develop and improve the statistics, additional analyses and evaluation of the different data sources are required.

9 References

Trafikanalys. ”Resvanor, Sverige”. Trafikanalys, u.å. www.trafa.se/transportmonster/RVU-Sverige/

Trafikanalys. ”PM 2023:6 Hur väl kan mobilnätsdata beskriva våra resvanor”, u.å. www.trafa.se/etiketter/transportovergripande/hur-val-kan-mobilnatsdata-beskriva-vara-resvanor-14193/

Trafikanalys. ”Regional linjetrafik och Körsträckor”. Trafikanalys, u.å. www.trafa.se/kollektivtrafik/kollektivtrafik/ and www.trafa.se/vagtrafik/korstrackor/

Zhang, L.-C. ”(2024) Disaggregation of trips using MNO data”. Opublicerat, MNO-MINDS.

SCB. ”Sveriges officiella statistik”. www.scb.se/om-scb/samordning-av-sveriges-officiella-statistik/

Sinno Jialin Pan and Qiang Yang Fellow, IEEE. ”(2009) A Survey on Transfer Learning”. www.cse.ust.hk/~qyang/Docs/2009/tkde_transfer_learning.pdf

James and Stein, Berkeley Symp. on Math. Statist. and Prob., 1961: 361-379 (1961). ”Estimation with Quadratic Loss”. https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fourth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Estimation-with-Quadratic-Loss/bsmsp/1200512173

Fay, R.E. and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to Census data. Journal of the American Statistical Association, Vol. 74, No. 366 (Jun., 1979), pp. 269-27.

Appendix – County List

Table 11. List of counties (alphabetical order).

Number	County
1	Blekinge County
2	Dalarnas County
3	Gotlands County
4	Gävleborgs County
5	Hallands County
6	Jämtlands County
7	Jönköpings County
8	Kalmar County
9	Kronobergs County
10	Norrbottens County
11	Skåne County
12	Stockholms County
13	Södermanlands County
14	Uppsala County
15	Värmlands County
16	Västerbottens County
17	Västernorrlands County
18	Västmanlands County
19	Västra Götalands County
20	Örebro County
21	Östergötlands County

Transport Analysis is a Swedish agency for transportpolicy analysis. We analyse and evaluate proposed and implemented measures within the sphere of transportpolicy. We are also responsible for official statistics in the transport and communication sectors. Transport Analysis was established in 2010 with its head office in Stockholm and a branch office in Östersund.

TRAFIK ANALYS

Transport Analysis
Rosenlundsgatan 54
SE-118 63 Stockholm

Phone +46 10 414 42 00
trafikanalys@trafa.se
www.trafa.se