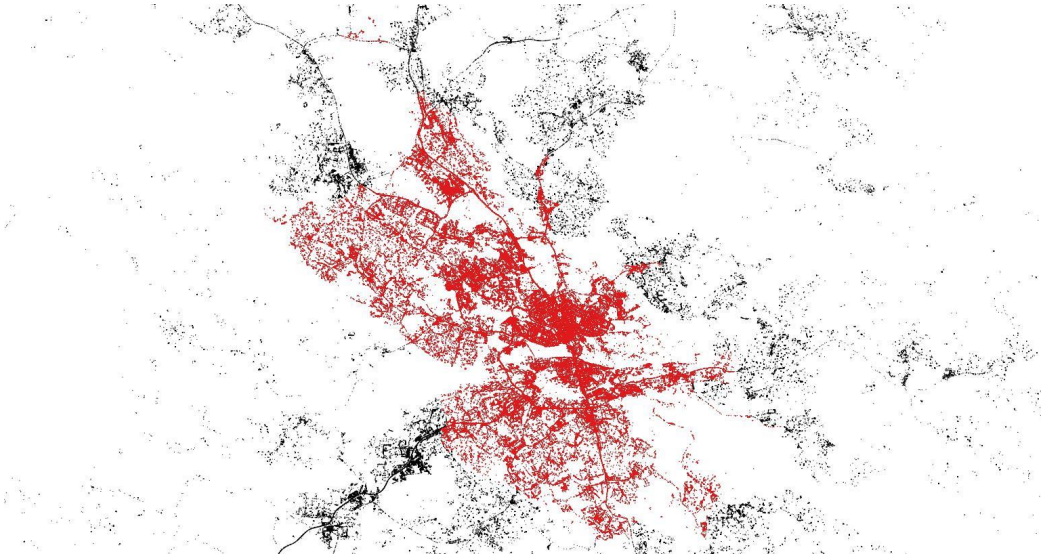


CODE PROBE

CONTINUOUS TRAVEL DEMAND AND LINK FLOW

ESTIMATION BASED ON GPS PROBE DATA

CODE PROBE
CONTINUOUS TRAVEL DEMAND AND LINK FLOW ESTIMATION BASED ON GPS PROBE DATA

Project title	**Continuous travel demand and link flow estimation based on GPS data (CODE PROBE)**
Swedish project title	**Probe-data för kontinuerlig skattning av OD-matriser och länkflöden**
Diarienummer	**TRV 2021/22404**
Authors	**Malin Andersson, Ramboll** **Joakim Ahlberg, Ramboll** **Marian Codreanu, LiU** **David Gundlegård, LiU** **Joachim Johansson, Ramboll** **Clas Rydergren, LiU** **Anders Sjöholm, Ramboll** **Susanna Sjöstrand, Ramboll** **Nikolaos Tsanakas, LiU** **Guang Wei, LiU**
Contact persons	**David Gundlegård (david.gundlegard@liu.se)** **Anders Sjöholm (anders.sjoholm@ramboll.se)**
Recipients	**Trafikverket** **Project sponsor Fredric Almkvist** **Administrative officials Per Eriksson, Matilda Mårtensson**
Version	**1.1**
Datum	**2024-01-02**

SAMMANFATTNING

Nya storskaliga datakällor för mobilitet, som GPS-baserad probe-data och mobilnätdata, har på senare tid fått ökat intresse inom trafikplaneringsapplikationer [1], [2]. Dessa datakällor är baserade på kontinuerliga positionsobservationer av ett urval av enheter som möjliggör en helt ny förståelse av spatiotemporala mobilitetsmönster både inom och mellan städer. En storskalig mobilitetskälla av särskilt intresse för trafikplanering är GPS-baserad probe-data, som inkluderar högupplösta mobilitetsdata för några procent av fordonsparken. Datakällan har potential att automatisera trafikstatistik på nätverksnivå för tillämpning inom både trafikprognoser och nationell ekonomisk analys. Trafikverket har redan anskaffat aggregerade restider baserat på dessa probe-data och i förstudien ”Förstudie kring användning av probedata för skattning av tidsvarierande reseefterfrågan och trafiktillstånd” [3] har egenskaperna hos rådata och möjliga applikationsområden analyserats för svenska förhållanden. Förstudien identifierade länkflödesskattningar som en prioriterad datatyp som är användbar för trafik- och infrastrukturplanering. Nyligen har dock forskning också använt probe-data som indata till skattning av efterfrågan i form av OD-matriser [4].

Detta projekt syftar till att kombinera lokala probe-baserade länkflödesuppskattningar med konceptet datadriven nätverksutläggning (DDNA) och skattning av OD-matriser [5], som nyligen utvecklats vid LiU, för kontinuerlig (24/7) uppskattning av efterfrågan och länkflöde. Den datadrivna nätverksutläggningen kopplar ihop efterfrågan med lokala länkflödesuppskattningar och möjliggör konsistent skattning av både OD-matriser och länkflöde kontinuerligt över tid, baserat på högupplöst probe-data som indata.

Resultaten visar att detaljerad probe-data kan användas för 1) direkt länkflödesuppskattning, 2) identifiering av ovanliga händelser i trafik och 3) uppskattning av en fördelningsmatris för mappning mellan OD och länkflöden. Metoderna för att prediktera länkflödesmätningar med hjälp av probe-data kan förklara en stor del av variationen i data på timnivå, men felen är relativt stora, speciellt för länkar med få observationer. Metoder baserade på maskininlärning kan förbättra predikteringens noggrannhet, men ytterligare arbete behövs för att analysera vilken kombination av förklaringsvariabler som bör användas och vilken typ av modell som ger den bästa avvägningen mellan noggrannhet och varians vid prediktering av länkflöden även för länkar utan länkflödesmätningar.

Direkt skalning av OD-flöden är utmanande för kontinuerlig uppskattning av OD-flöden på timnivå, om inte den rumsliga aggregeringen är mycket stor. Genom att aggregera data för längre tidsperioder är det dock möjligt att få rimliga OD-flöden även från GPS-data, åtminstone när de aggregeras till zonflöden.

Probe-data kan användas för att detektera händelser i trafiksystemet och att använda storskaliga mobilitetsmönster kan förbättra detekteringen jämfört med att endast använda lokala skattningar av länkflöden. Händelser med begränsad tidsutbredning som inträffar på vägar med små flöden utan större förändringar i storskaliga rörlighetsmönster är dock svåra att upptäcka med endast probe-data.

Att uppskatta en fördelningsmatris med hjälp av probe-data möjliggör konsistent och kontinuerlig OD- och länkflödesuppskattning. Noggrannheten i länkflödesskattningarna baserade på DDNA liknar noggrannheten vid användning av lokala skalningsmetoder baserade på maskininlärning, med fördelen att vi får en förståelse för OD- och ruttflöden samtidigt. Att uppskatta fördelningsmatrisen baserat på probe-data är utmanande och mer arbete krävs för att utvärdera både fördelningsmatrisen, länkflödena och OD-flödena för längre tidsperioder och nya geografiska områden.

SUMMARY

New emerging large-scale mobility data sources like GPS probe and mobile network data have recently gained increasing interest for traffic planning applications [1], [2]. These data sources are based on continuous location observations of a sample of devices that enable a completely new understanding of spatiotemporal mobility patterns both within and between cities.

A large-scale mobility source of special interest for traffic planning is GPS probe data, that includes high resolution mobility data for a few percent of the vehicle fleet. The data source has potential to automate network-wide traffic statistics for both traffic forecasts and national economic analysis. Trafikverket are already procuring travel time aggregates based on these GPS probe data and in the pre-study "Förstudie kring användning av probedata för skattning av tidsvarierande reseefterfrågan och trafiktillstånd" [3], characteristics of the raw data and possible application areas has been analyzed for Swedish conditions. The pre-study identified link flow estimates as a prioritized data type that is useful for traffic and infrastructure planning. However, recently, new research has also utilized GPS probe data as input to origin-destination travel demand estimation [4].

This project aims at combining local probe-based link flow estimates with the concept of data-driven network assignment and travel demand estimation [5], recently developed in LiU, for continuous (24/7) estimation of travel demand and link flow. The data-driven network assignment connects the travel demand with local link flow estimates and enables consistent estimation of both travel demand and link flow over time, based on raw GPS probe data as input.

The results show that detailed GPS probe data can be used for 1) direct link flow estimation, 2) identifying unusual events in traffic and 3) estimating an assignment matrix for mapping between OD and link flows.

The methods for predicting link flow measurements using GPS probe data can explain a large part of the variation in the data in hourly level, but the errors are relatively large, especially for links with few observations. Methods based on machine learning can improve the link flow prediction accuracy, but further work is needed to analyse which combination of features that should be used and which type of model that will give the best bias-variance trade-off when predicting link flows also for links without link flow measurements.

Direct scaling of OD flows is challenging for continuous hourly estimation, unless the spatial aggregation is very large. However, by aggregating data for longer time periods it is possible to get reasonable OD flows also from GPS probe data, at least when aggregated to zone flows.

GPS probe data can be used to detect events in the traffic system and utilising large-scale mobility patterns can improve the detection compared to only using local estimates of link counts. However, events with limited extension in time that occurs on roads with small flows without major changes in large-scale mobility patterns are hard to detect using only GPS probe data.

Estimating an assignment matrix using GPS data enables consistent and continuous OD and link flow estimation. The accuracy of the link flow estimates for DDNA-based link flow estimates are similar to the accuracy when using local scaling methods based on machine learning, with the advantage that we get an understanding of OD and route flows at the same time. Estimating the assignment matrix based on GPS probe data is challenging and more work is needed to evaluate both the assignment matrix, the link flows and the OD flows for larger time periods and new geographical areas.

INNEHÅLLSFÖRTECKNING

1. INTRODUCTION

1.1 Background

New large-scale data sources for mobility, such as GPS-based probe data and mobile network data, have recently gained increased interest in traffic planning applications [1], [2]. These data sources are based on continuous position observations of a selection of devices that enable a completely new understanding of spatiotemporal mobility patterns both within and between cities.

A large-scale mobility source of particular interest for traffic planning is GPS-based probe data, which includes high-resolution mobility data for a few percent of the vehicle fleet. The data source has the potential to automate traffic statistics at the network level for application in both traffic forecasting and national economic analysis. The Swedish Transport Administration has already acquired aggregated travel times based on this probe data and in the preliminary study "Pre-study on the use of probe data for estimating time-varying travel demand and traffic conditions" [3] the characteristics of the raw data and possible areas of application have been analyzed for Swedish conditions. The feasibility study identified link flow estimates as a priority data type useful for traffic and infrastructure planning. Recently, however, research has also used probe data as input to demand estimation in the form of OD matrices [4].

Within the project, local probe-based link flow estimations is combined with the concept of data-driven network assignment and demand estimation [5], [6], recently developed at LiU, for continuous (24/7) estimation of demand and link flows. The data-driven network assignment connects demand with local link flow estimates and enables consistent estimation of both demand and link flow over time, based on detailed probe data as input.

1.2 Purpose

The project aims to develop methods to continuously estimate link flows and OD matrices in transport networks using detailed probe data from GPS-equipped vehicles. Furthermore, the project also aims at identifying and demonstrating the potential for various applications of continuous estimation of OD matrices and link flows in traffic planning.

1.3 Methodology

To identify the most important use cases and understand how GPS probe data is used today, the project started with a survey that was sent to potential users of the project result.

Two different approaches was used for estimating link flows, one local estimation based on machine learning and one network-wide estimation based on data-driven network assignment. In the machine learning approach training data for selected link flow sensors are used to estimate models that use GPS probe samples to predict total link flows. In the network-wide approach an assignment matrix that maps OD flows to link flows is first estimated and then used for estimating both an OD matrix and link flows. A method for identifying outlier mobility patterns based on detailed GPS probe data is also developed.

The methods were evaluated using a five-week dataset with GPS-probe data, mobile network data, traffic model data (SAMPERS) and link count data in Stockholm.

2. USE CASE IDENTIFICATION

To identify the exact case studies, first, we performed a study on the current use of GPS data by different transportation planning-related actors in Sweden. The purpose of this study was to find out what needs exist among traffic analysts, planners and investigators, and how methods for GPS data analysis can be designed to meet those needs.

For that we developed a web-based survey. The survey included an introduction briefly explaining the concept and potential applications of GPS data in transportation. The introduction also articulated the main goal of the project, emphasizing the development of methods for estimating origin-destination matrices, route choice and link flows continuously over time, based on high-resolution probe data as input. The rest of the survey consisted of questions regarding familiarity with the GPS data and experience using it, as well as opinion on barriers deterring the authorities from using the GPS data and open question for possible data applications for the existing projects.

As a result, we collected 30 answers. Out of which, 14 answers were from the communes' authorities, 14 from Trafikverket, one from a regional authority and one from "other" type of authorities / organizations. Two thirds of the respondents were traffic analysts and traffic planners.

The answers showed minor experience of the respondents in GPS data applications. Only 25% of the respondents had some experience in using GPS data in their work, the rest had either no experience or very little. None of the respondents stated having substantial experience in GPS data use. Traffic studies and pre-evaluation studies were among the most common applications, followed by flow- and speed-maps.

The questions related to the potential of the use of the GPS data, on the contrary to the respondents' minor experience, showed substantial interest in the topic. None of the respondents answered negatively to a question on the existence of interest and wish to use the GPS data in traffic analysis and planning. The respondents voted on different potential applications of the GPS data in their work. Most votes gained traffic studies (26 votes), planning in early stages, for example research/preliminary studies (23 votes), traffic flow maps (23 votes) and commuter pattern maps (21 votes). Speed maps appeared to be interesting only to a half of the respondents (15 votes). The main barriers hindering the use of GPS data, listed by the respondents, included lack of developed methods, missing processes and routines for the GPS data use and the data monetary costs.

One third of the respondents provided suggestions regarding case studies where the GPS data could be potentially applied for an existing or planned project. Most of those respondents were interested in calculation of link flows, half of them were interested also in route choice analysis, and only two respondents wished to estimate origin-destination matrices. Most of the proposed case studies covered projects related to both passenger and freight transport (e.g., construction of a new road or closure of an existing one). The projects were diverse in terms of geographic location and coverage. Half of the projects were for non-urban transport and another half for inter-city one.

The respondents, who provided the suggestions of the case studies and agreed to be contacted, received an invitation to participate in a meeting dedicated to their proposal and other potential needs. Four of those confirmed their availability for a meeting. Those four included transport specialists from three communes, including Stockholm, and representatives from a local office of Trafikverket. Only one of the four respondents had substantial experience in use of the GPS data.

An individual meeting was held with each of the four respondents, where we briefly presented the project and the respondents shared their views, concerns, and questions on the GPS data use, as well as suggested potential case studies.

Table 1 Summary of the case studies discussed with the survey respondents during the follow-up meetings

Organisation	Potential case studies	Type of GPS data application
Trafikverket, Region Nord	Study travel times and link flows on road 66 between Malung and Scandinavian Mountains Airport, including rural roads, to understand how the traffic is distributed.	Link flows (and travel times) and route choice
Ystad	- Linking GPS data to previous work with traffic strategy carried out by Ramboll in 2019 – focus on current situation, speed reductions, variation in volumes; – A similar / follow-up study within a few years; - Connection of GPS data to the Trafikverket's ÅVS E65; - Linking to the investigation of Dragongatan and new construction - traffic generation and load - Need tools / model to look at how expansions affect the flows; - Evaluation of indicators from the traffic strategy (traffic jams, etc.), following up indicators the with RVU.	OD matrices and link flows
Stockholm Stad	- Analysis of link flows along bridges in Stockholm; - Noise mapping – based on real speed data – and estimating flows; - Traffic models – find suitable places for measurements; - Maps – description Stockholm traffic system – hotspots for congestion, etc. – historical data.	OD matrices and link flows
Nyköping	- Accessibility to Nyköping travel center, construction starts in 2023, analysis of through traffic; - Brunnsgatan will be closed off to change railway bridges: analysis of traffic during closures is of interest - Strengthening the entrance to the E4, current situation analysis for route selection over the bridge – finding alternative routes for these trips; - Map, how to redirect traffic during shutdowns. - Construction of roundabout in the southern part of the city, before/after study. How did it affect traffic to and from the city (especially, heavy traffic); - Map current/normal situation.	OD matrices and route choice

The discussion of the potentially interesting applications helped us to identify the case studies for the project and the authorities interested in these studies. However, the final choice of the studies had to be limited because of the data availability and its penetration rate. Data availability implies both time and geographical coverage. The INRIX GPS data available for this project includes data for the Stockholm Region and Norrköping, 6 weeks in 2019 (September 16-30, October 1-20), with Nyköping being also partially covered by the data. Another data limitation is

the data penetration rate, which is the number of observations for particular roads per unit of time. The data provides reasonably good number of observations for the bigger roads and very small sample for the smaller roads. For example, for the study of the link flows between Malung and Scandinavian Mountains Airport (listed in Table 1) smaller roads were of most interest to help to understand how the tourist flows are distributed in high season, but the penetration rate of the INRIX GPS data would not allow this kind of study.

With the above considerations, the choice set was reduced for two case studies: Nyköping and Stockholm city. For Stockholm city the time for which the data is available constrained the choice of the case studies. So, we had to find the exact events, for which we analyze the data, in the 6 weeks of September-October 2019, rather than use case studies proposed during the discussion with the city authorities. Nyköping was selected as the second case study, despite the relatively low INRIX data penetration rate, to check what kind of analysis could be done with the available data for a smaller city given the data limitations.

3. DATASETS

To evaluate the methods developed in the project, a five-week dataset with GPS-probe data, mobile network data, traffic model data (SAMPERS) and link count data was compiled for Stockholm. The time period for the data was 2019-09-16 to 2019-10-21.

3.1 GPS probe data

A detailed description of the detailed GPS probe data from INRIX that is used in this project can be found in [6]. The Stockholm dataset contains 421 000 trips that passes by the chosen area, where 323 000 trips start inside the area and 324 000 trips end in the area. Figure 1 shows the spatial distribution of trips in the area and Figure 2 shows the temporal distribution of the GPS probe data.

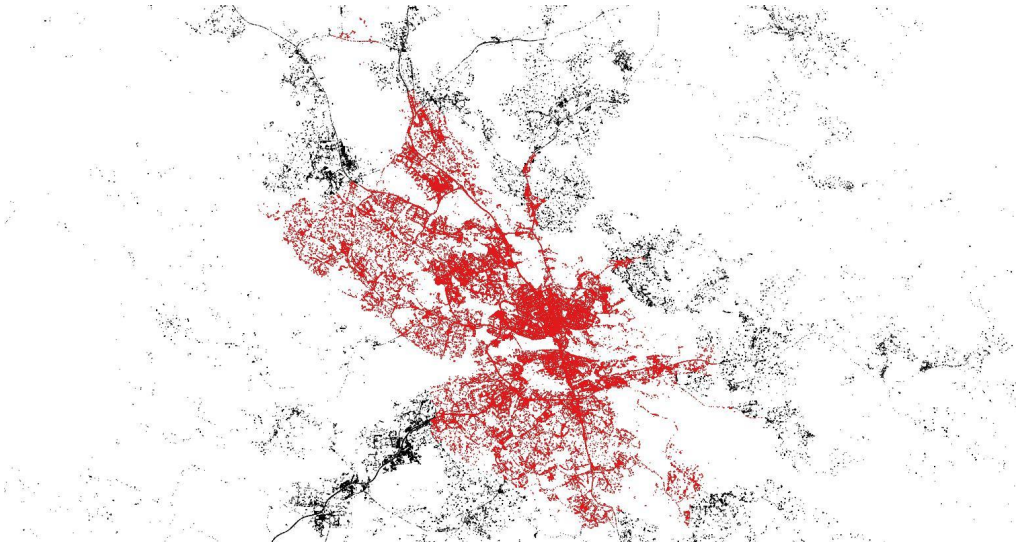


Figure 1 Start points of trips starting inside the area (red) together with start points for nearby trips (black)

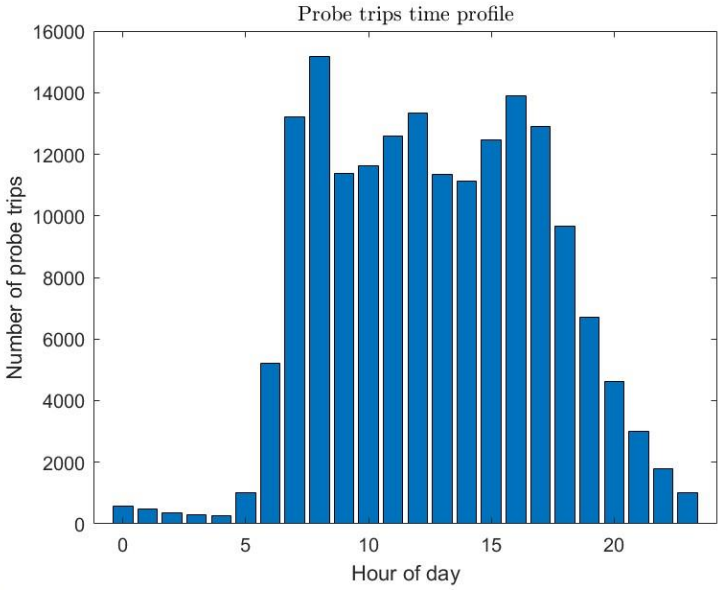


Figure 2 Temporal profile of probe trips start time during weekdays.

3.2 Road network

The road network for the area includes just above 83 000 links, where approximately 5000 links correspond to highways, major arteries and major roads, and the remaining are neighborhood streets according to the INRIX functional road class. The road network is illustrated in Figure 3.

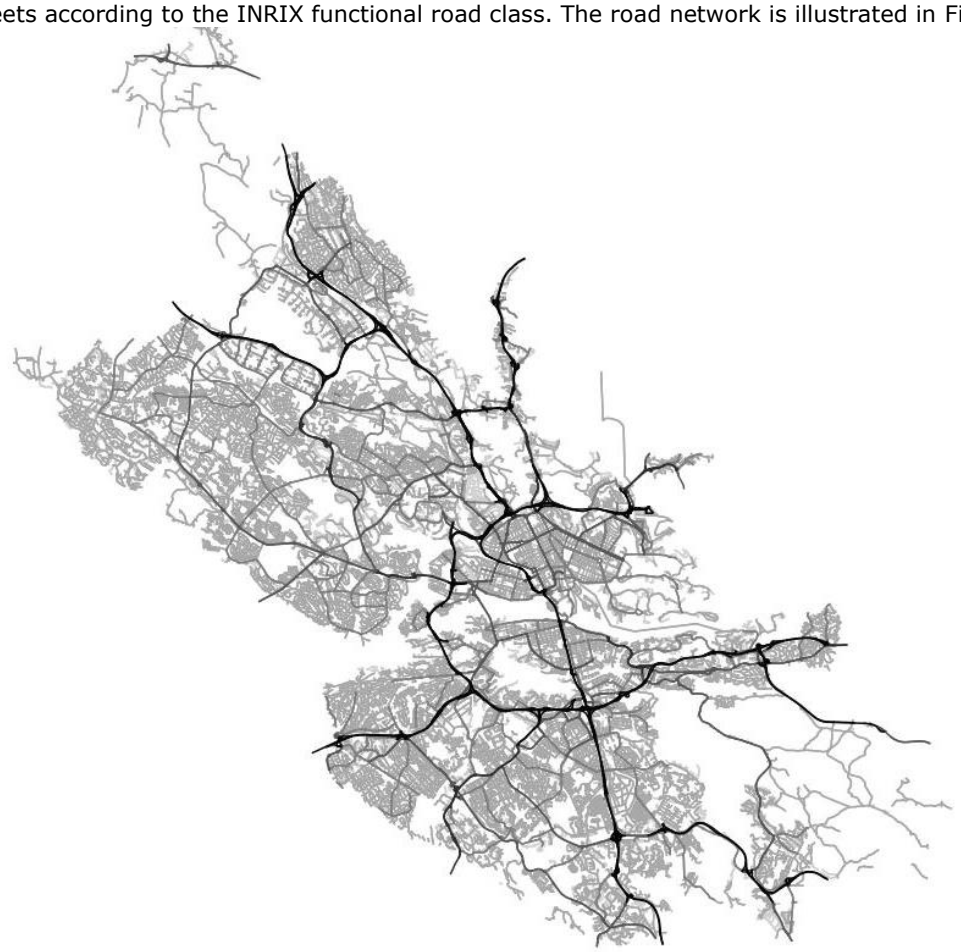


Figure 3 Overview of the road network in the test area with a total of 83 366 links.

3.3 Link flow observations

An important part of the project is to compare estimated link counts with link flow observations. The link flow observation dataset contains data from both MCS radar stations and temporary link counts. Totally 220 000 link count observations from 700 different sensors are available. The locations of the sensors are shown in Figure 4.

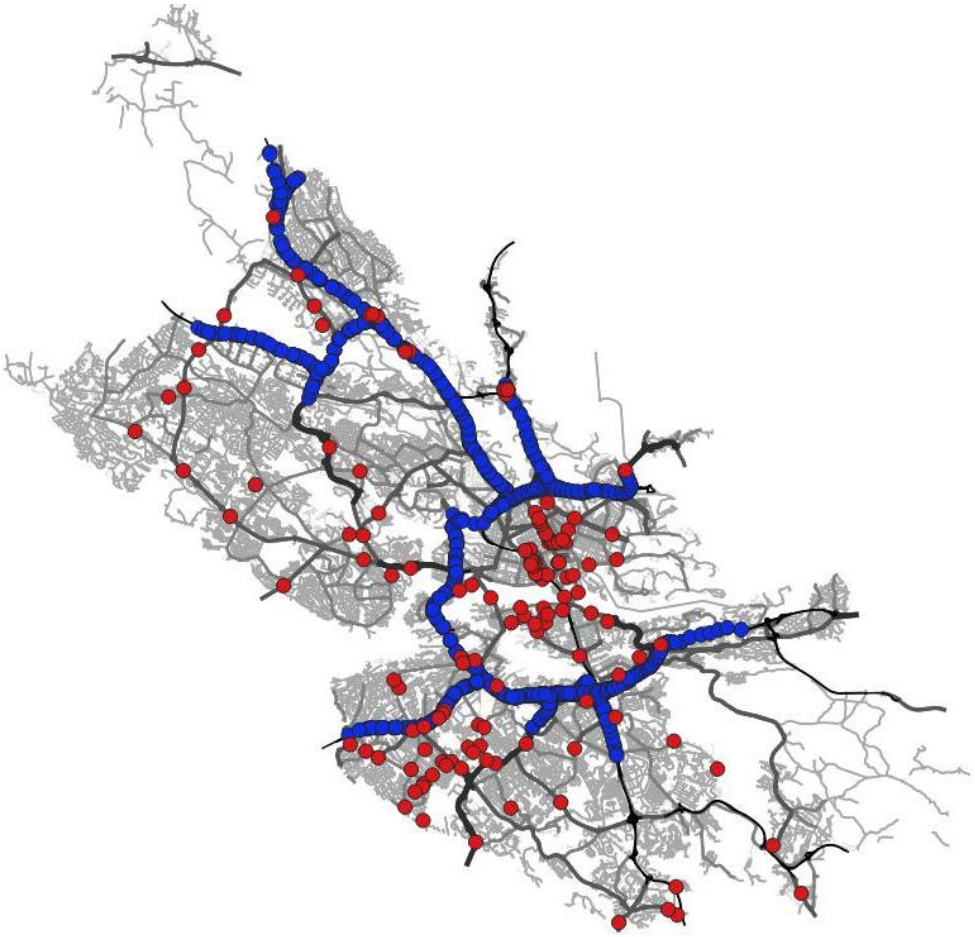


Figure 4 Link flow measurements from MCS (blue) and temporary link counts (red).

3.4 Mobile network data

Anonymized and aggregated mobile network data is provided by the mobile operator Telia through their service Telia Crowd Insights [7]. The observed all-mode (walking, biking, micro-mobility, shared vehicles, car, trains, and public transport) travel data that is used is scaled to represent the full population. A more detailed description of the type of data used is available in [8]. Crowd Insights data is available in many different spatial resolutions, we have chosen the OD zoning illustrated in Figure 5 with 31 zones in the area. In this project we have mainly used the routed OD matrix classified as "ROAD" traffic in Crowd Insights. The temporal profile of the trips from mobile network data is shown in Figure 6.

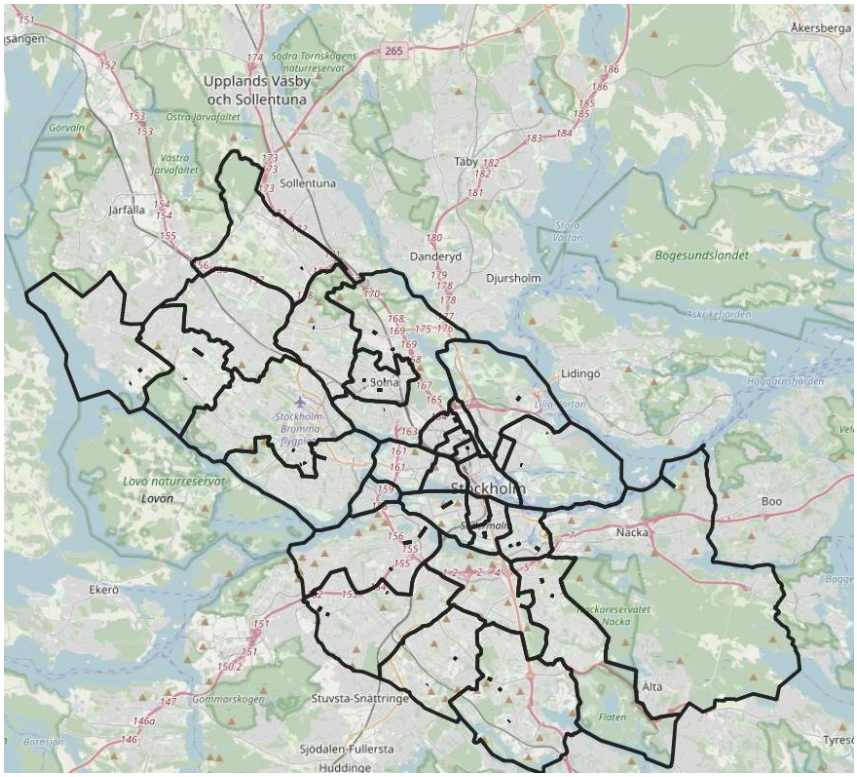


Figure 5 Telia zoning with 31 zones in the area

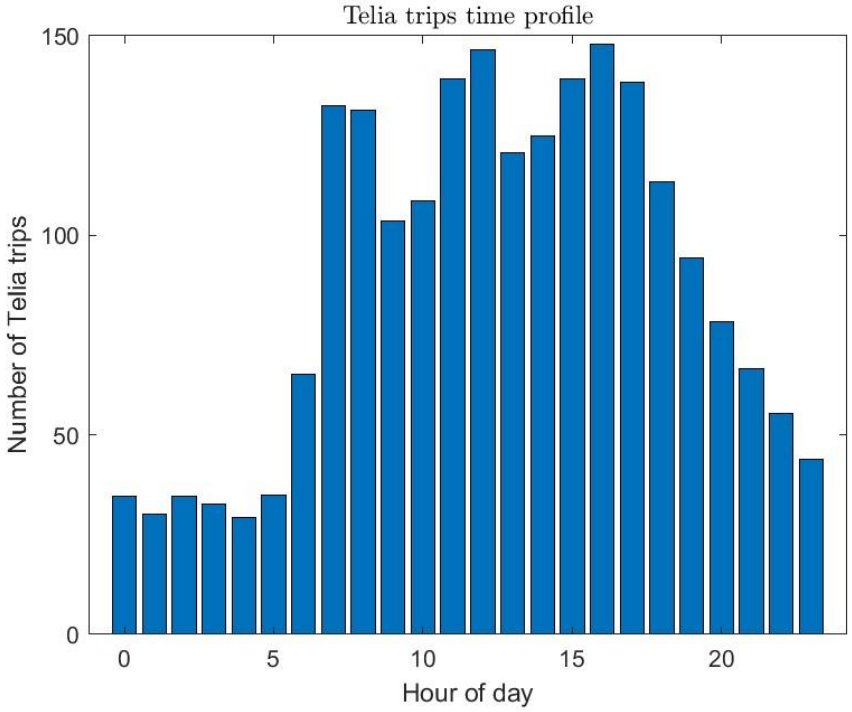


Figure 6 Average Temporal profile of the Telia data for weekdays during one week.

3.5 Demand model data

An OD matrix based on the demand model SAMPERS is used in the project. The OD matrix is for an average winter weekday for year 2020. The data includes predictions for morning peak, afternoon peak, low traffic and night traffic. The 558 zones in the area are illustrated in Figure 7 and the temporal profile of the OD matrix is shown in Figure 8.

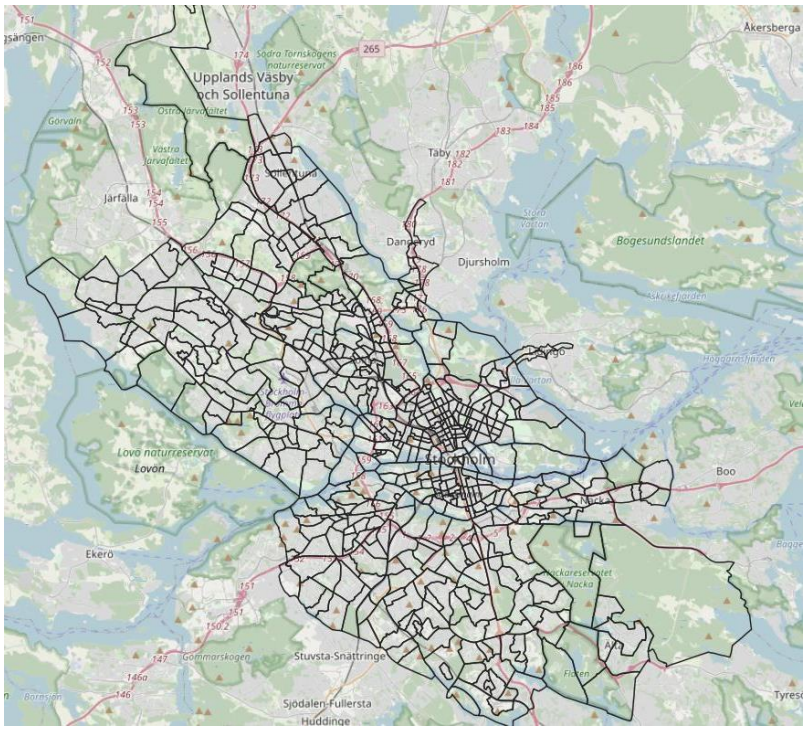


Figure 7 SAMPERS zoning with 558 zones in the area

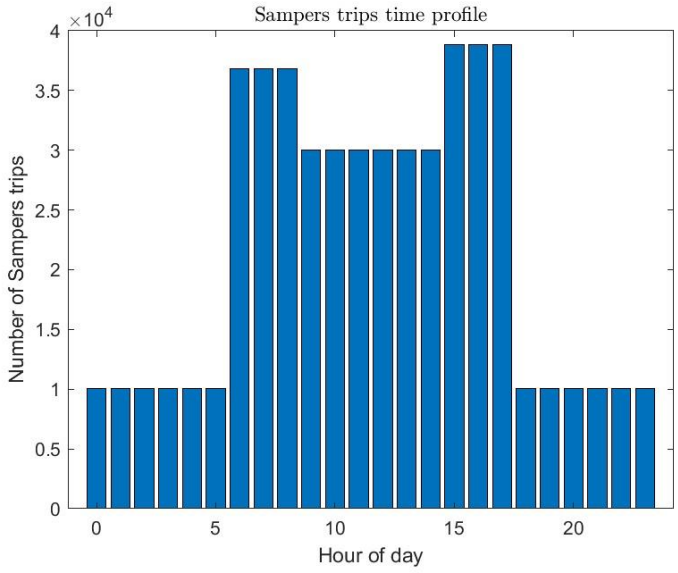


Figure 8 SAMPERS OD matrix time profile.

4. PROCESSING AND ANALYSIS OF GPS PROBE DATA

This chapter describes direct processing of detailed GPS probe data for the purpose of link flow estimation, OD estimation and major event detection. The methods described in this chapter focus on using the GPS probe data directly, without combining different types of data. However, for evaluation of the methods, and in some cases training of Machine learning models, we use other data sources like link counts, OD matrix data and incident data.

4.1 Link flow estimation

In this section we focus on estimating hourly link flows in the test area based on local scaling of the observed GPS probe data.

4.1.1 Dataset characteristics

Figure 9 to Figure 11 shows the average number of probe counts in the test area.

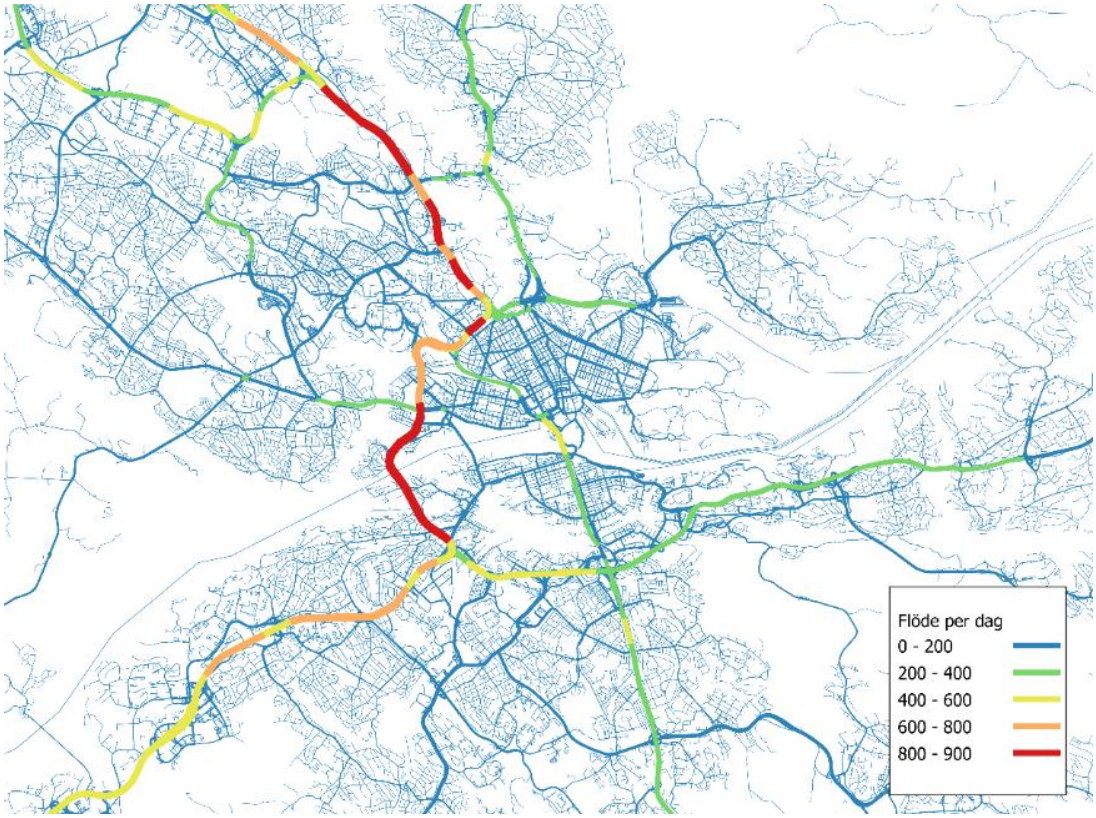


Figure 9 Average probe flow per day in the test area

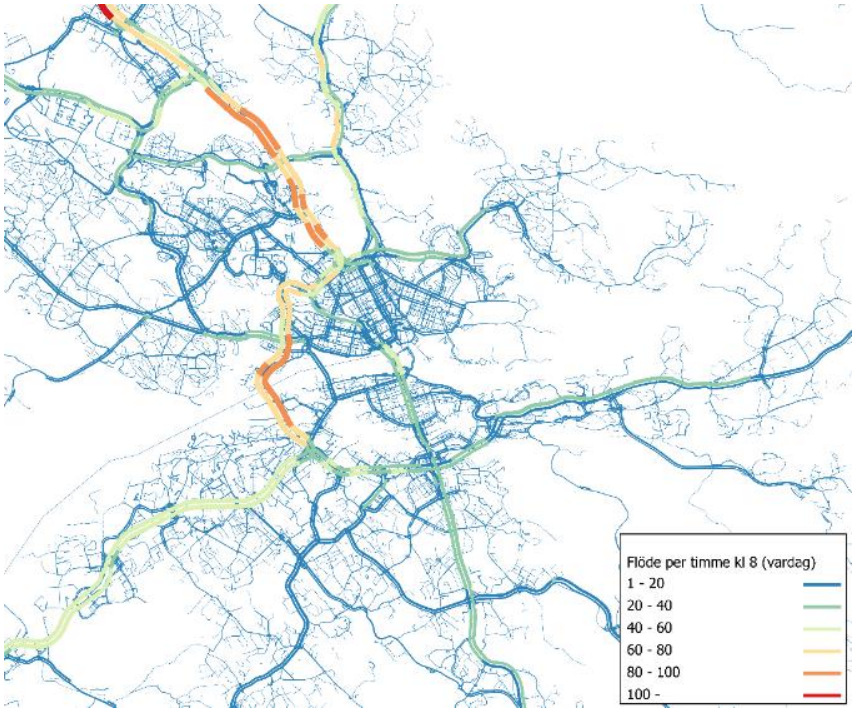


Figure 10 Average flow per hour at 8 am weekdays

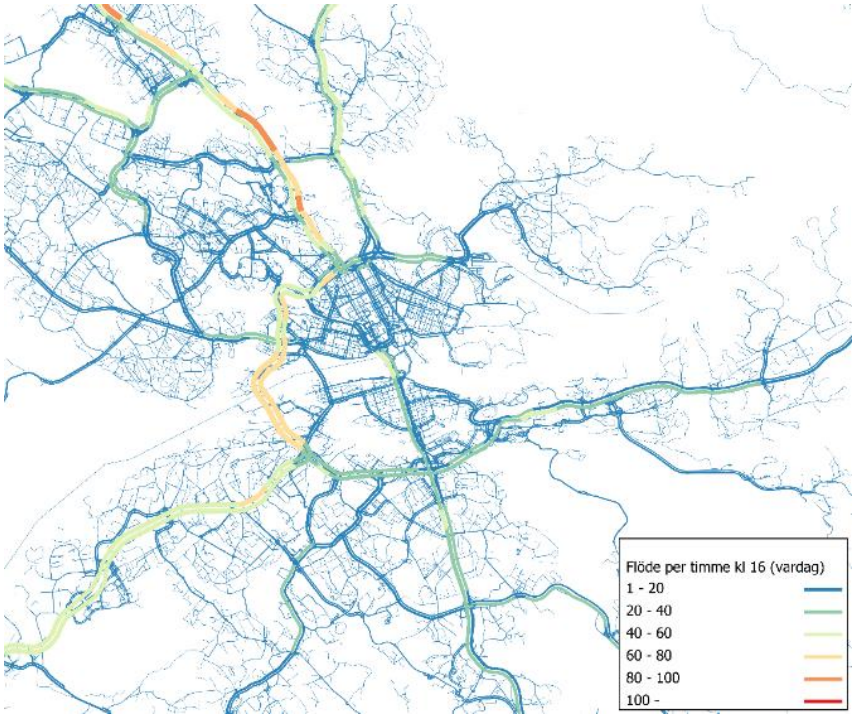


Figure 11 average flow per hour at 4 pm weekdays

The links with flow measurements is illustrated in Figure 12. An important characteristic of the GPS probe data is how the probe fraction (penetration rate) varies in space and time, which is illustrated in Figure 13 to Figure 19.

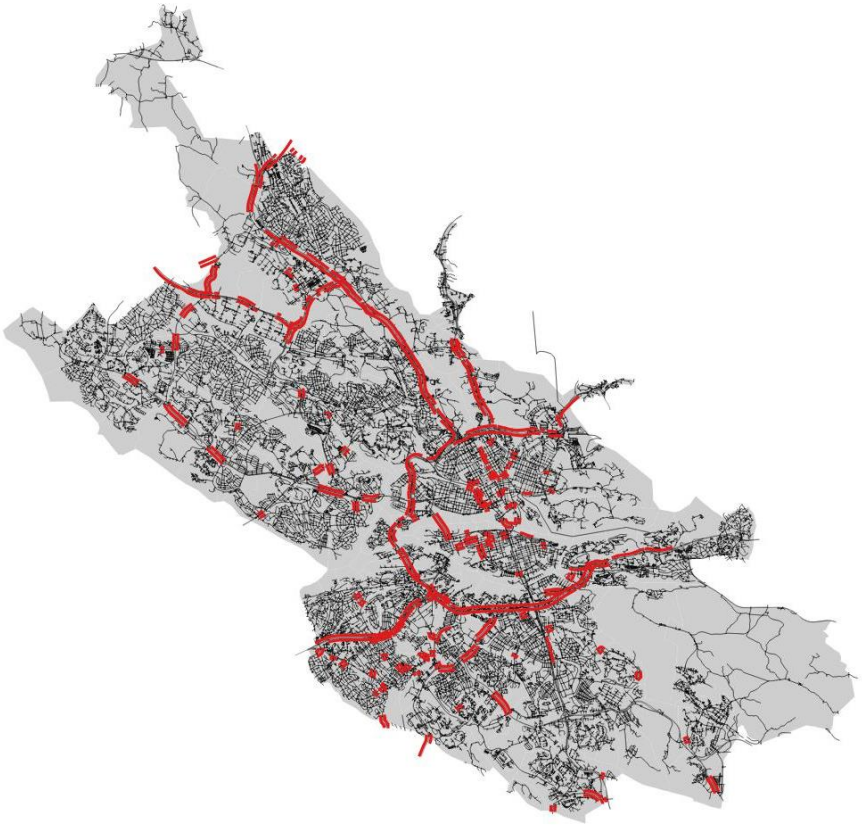


Figure 12 Links with flow sensor data.

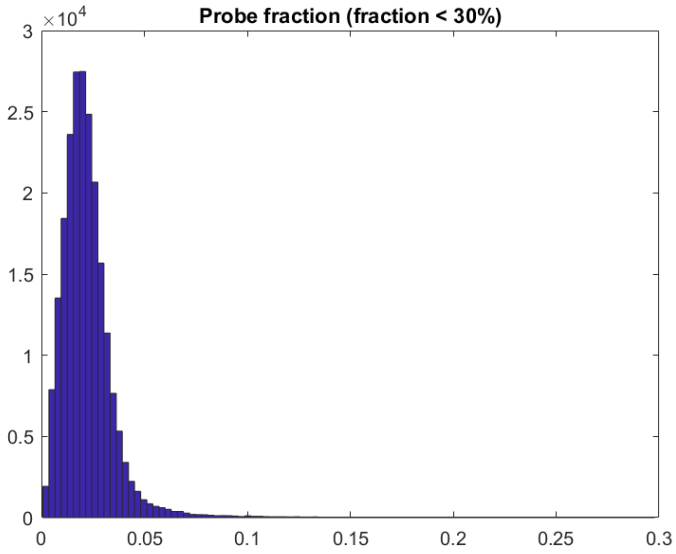


Figure 13 Probe fraction distribution.

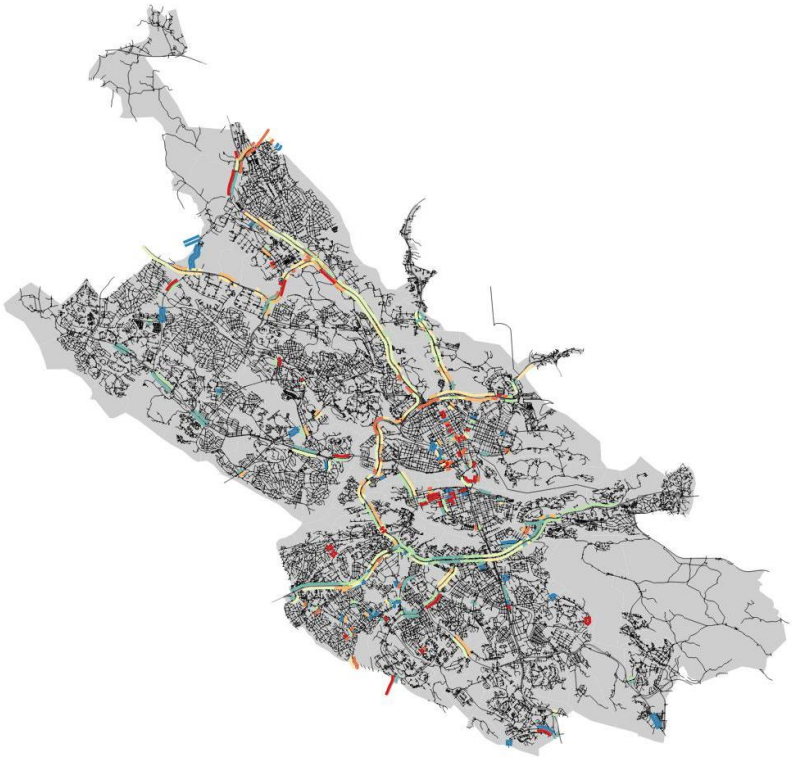


Figure 14 Average probe fraction per sensor

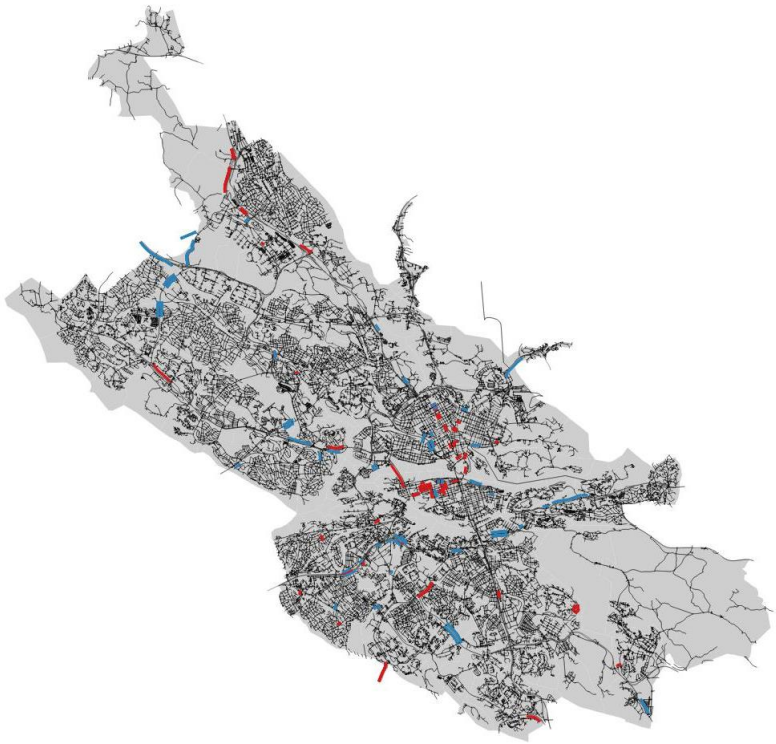


Figure 15 Hourly outlier probe fractions (blue < 0.002 red > 0.2)

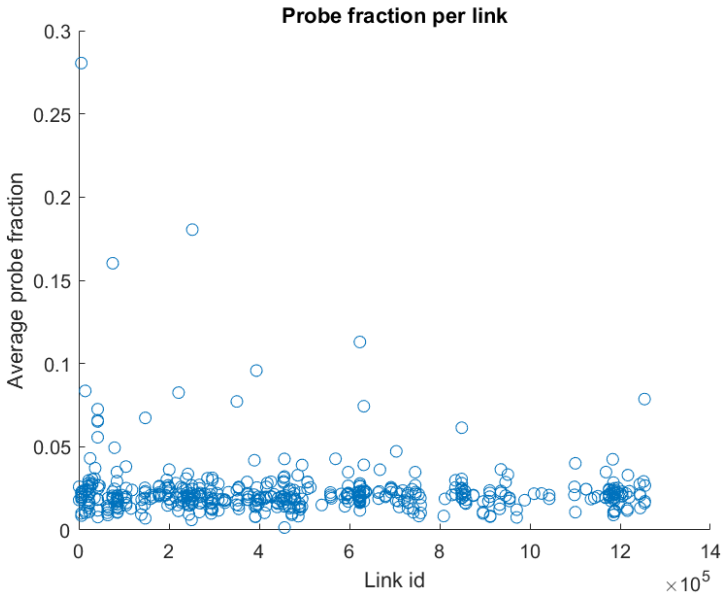


Figure 16 Probe fraction per link.

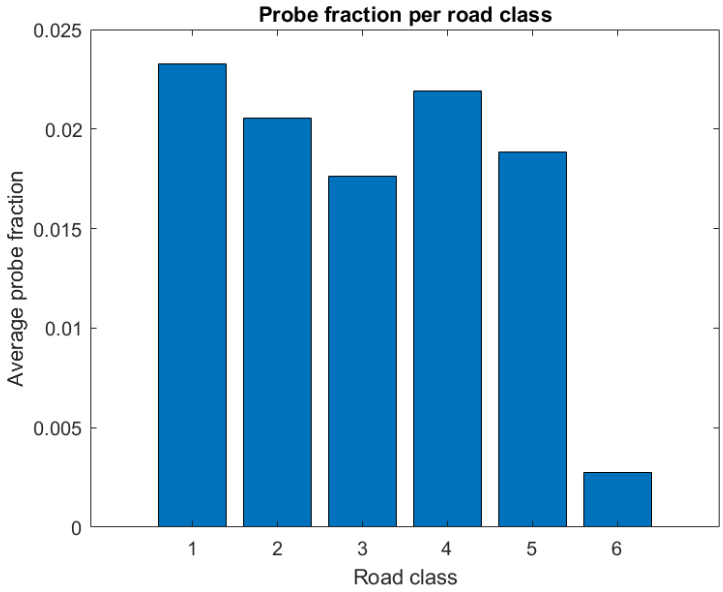


Figure 17 Probe fraction per road class.

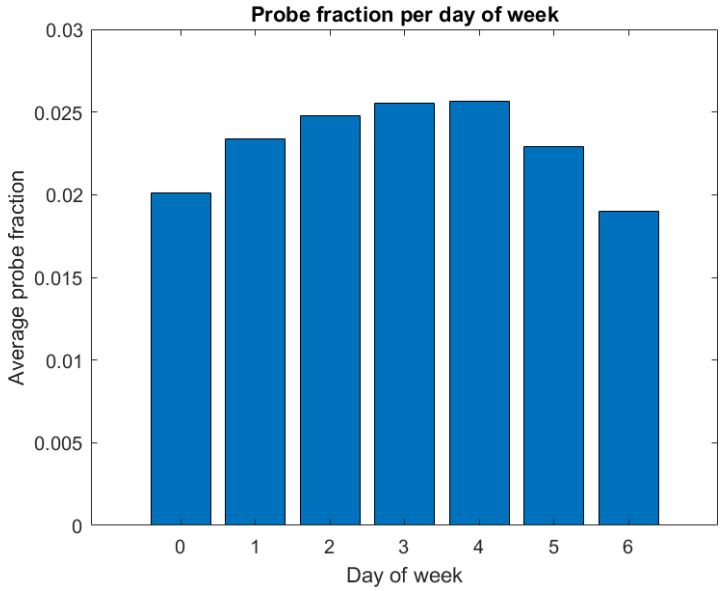


Figure 18 Probe fraction per day of week

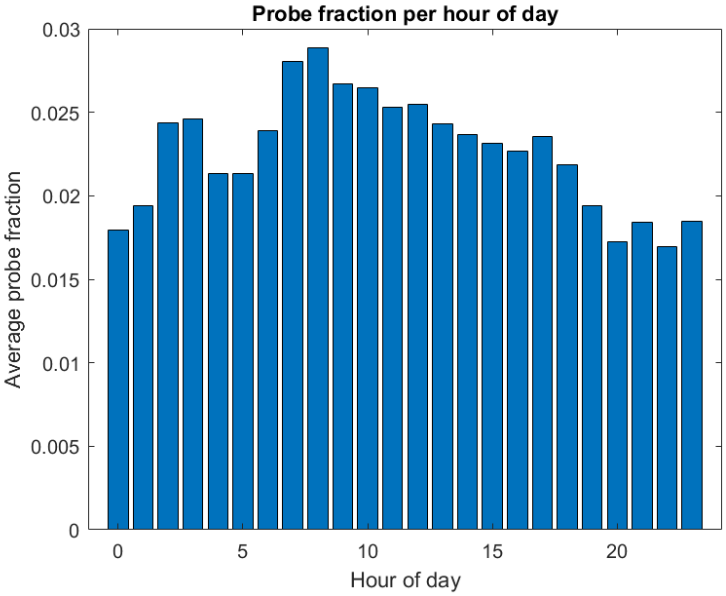


Figure 19 Probe fraction per hour of day

Figure 20 shows a scatter plot of probe counts and measured sensor flow for all sensors.

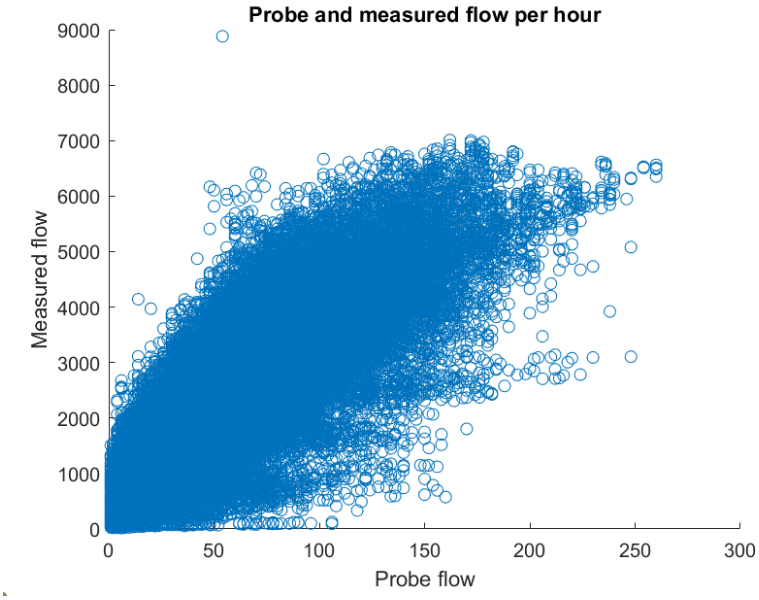


Figure 20 Scatter plot of probe flow VS measured flow for all sensors

4.1.2 Machine learning approach

To analyse the potential of using machine learning models to estimate link flows a dataset with eight features and one response variable is compiled. A description of the features and the response variable is presented in Table 2. The dataset contains 220 000 observations for 700 different sensors.

Five different methods for estimating link flows are evaluated. The first method is a basic benchmark method based on simple scaling of GPS probe data (called GPS fraction). Four different regression methods are also evaluated, these are Least Square, Poisson, L1 and weighted L1 regression. The features used in the models are *road_class*, *dow*, *hod* and *probe_flow*. A more detailed description of the different methods is available in the report appendix.

Table 2 Features and response variable of the link flow dataset

Feature	**Description**
segment_row_id	Identity of the link
road_class	Type of road according to openstreetmap, 1 = highway, 6 = small residential road
sensor_id	Identity of the sensor
sensor_type	Type of sensor, medy = temporary link count, mcs = fixed radar station
timestamp	Time of the measurement
dow	Day of week 0 = sunday, 6 = saturday, extracted from timestamp
hod	Hour of day, extracted from timestamp
probe_flow	Number of GPS probe vehicles that have passed the link during one hour
Response variable	**Description**
measured_flow	Total number of vehicles per hour measured by the sensor

The methods are trained and evaluated by randomly dividing the dataset into train and test data, where 80% is used for training and 10% is used for test. The methods are evaluated using the following metrics:

1) Root Mean Square Error, RMSE:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2}{N}}$$

2) Mean Absolute Error, MAE:

$$MAE = \frac{\sum_{i=1}^{N}\left|Y_i - \hat{Y}_i\right|}{N}$$

3) Coefficient of determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

4) MAPE:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|Y_i - \hat{Y}_i\right|}{Y_i}$$

The results for the different methods are shown in Figure 21 to Figure 24.

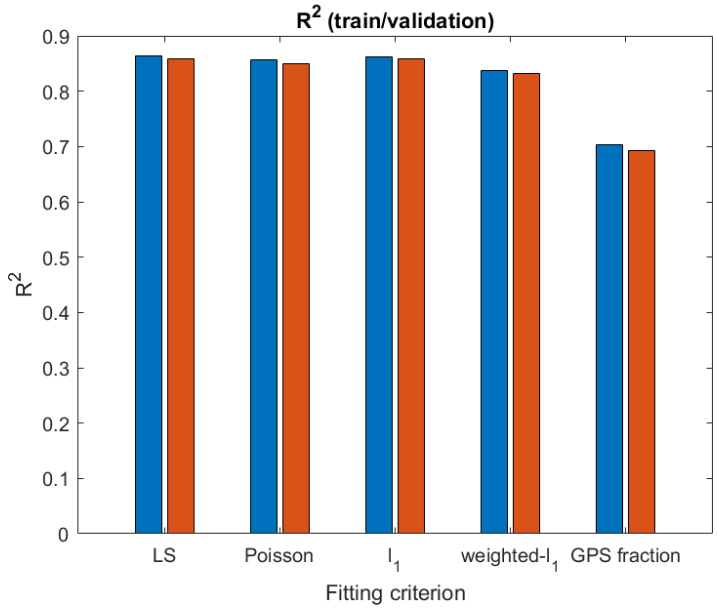


Figure 21 Coefficient of determination (R^2) for the different methods. Blue bars for training set error and red bars for test set error.

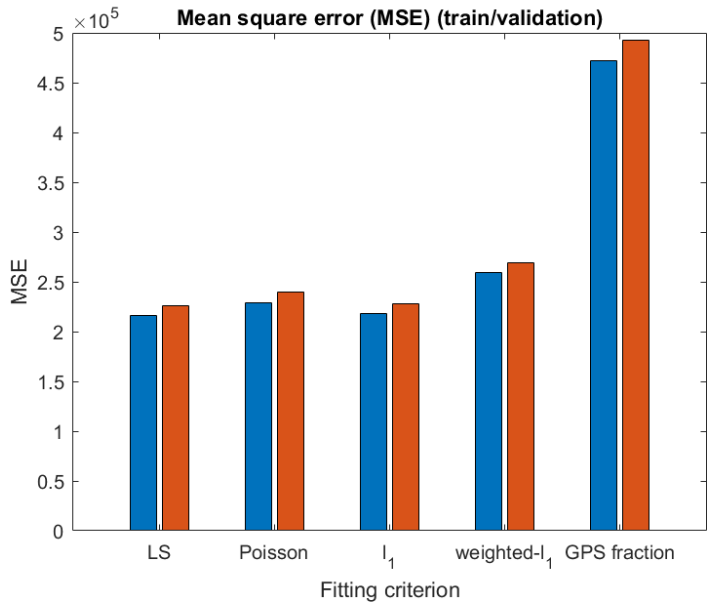


Figure 22 MSE for the different methods. Blue bars for training set error and red bars for test set error.

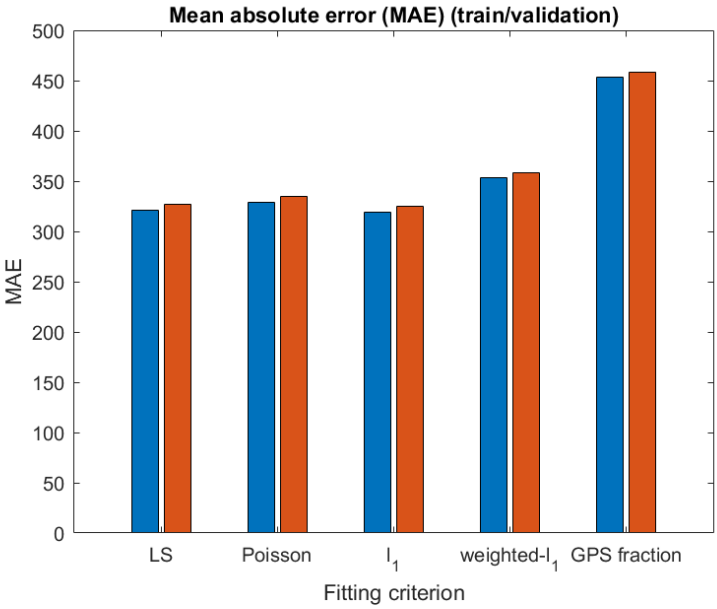


Figure 23 MAE for the different methods. Blue bars for training set error and red bars for test set error.

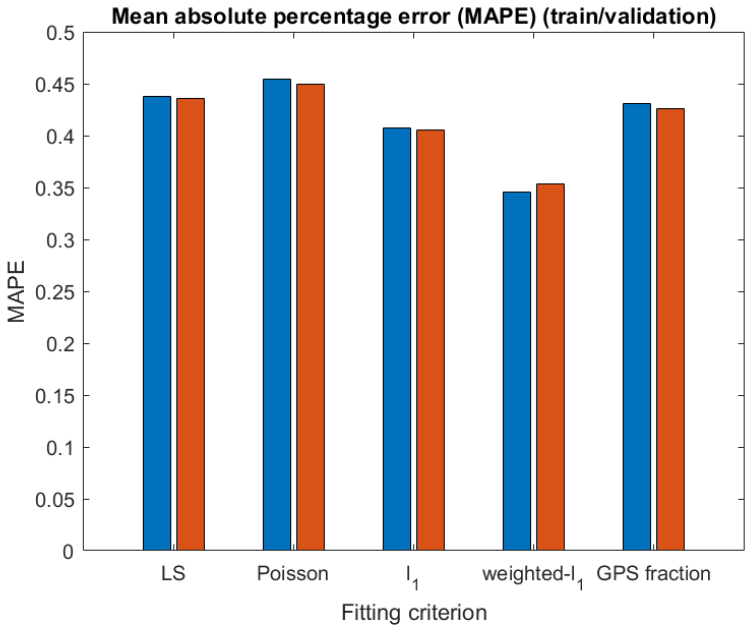


Figure 24 MAPE for the different methods. Blue bars for training set error and red bars for test set error.

The results show that a large part of the variance in the data is explained by the models, where the machine leraning models perform better than the simple GPS fraction model for all performance metrics. The results also show that the different machine learning methods perform clearly better when the model estimation method is tailored against the specific performance metric.

It should be noted that these results are obtained using only generic spatial (road_class) and temporal (hod and dow) features except the probe_flow feature. This means that the results is easier to generalize to links without sensor data. If for example sensor_id is used as a feature, the prediction error will be smaller, but the generalization to new links is more difficult.

4.2 OD estimation

Both GPS probe data and mobile network data includes direct observations of OD flows, and in this section we analyse the OD flow characteristics of the two data sources and compare it to SAMPERS OD flows. Figure 25 shows the detailed SAMPERS zoning and the more aggregated Telia zoning used in the test area. The analysis in this section is focused on daily inter zone OD flows in the area. Intra zone and external zone OD flows are not included. For the Telia data, OD flows for weekdays 2019-09-16 to 2019-09-20 is used. For GPS probe data, all weekdays in the 5-week period are used.

Figure 26 to Figure 28 shows the spatial distribution of trip start points in the Telia zoning level. Figure 29 to Figure 31 shows scatterplots of the three different data sources against each other. Figure 32 to Figure 34 shows histograms of different data source ratios. Figure 35 to Figure 38 show scatter plots of the data sources for specific days.

Figure 25 Telia (thick lines) and SAMPERS (thin lines) zoning for the test area.

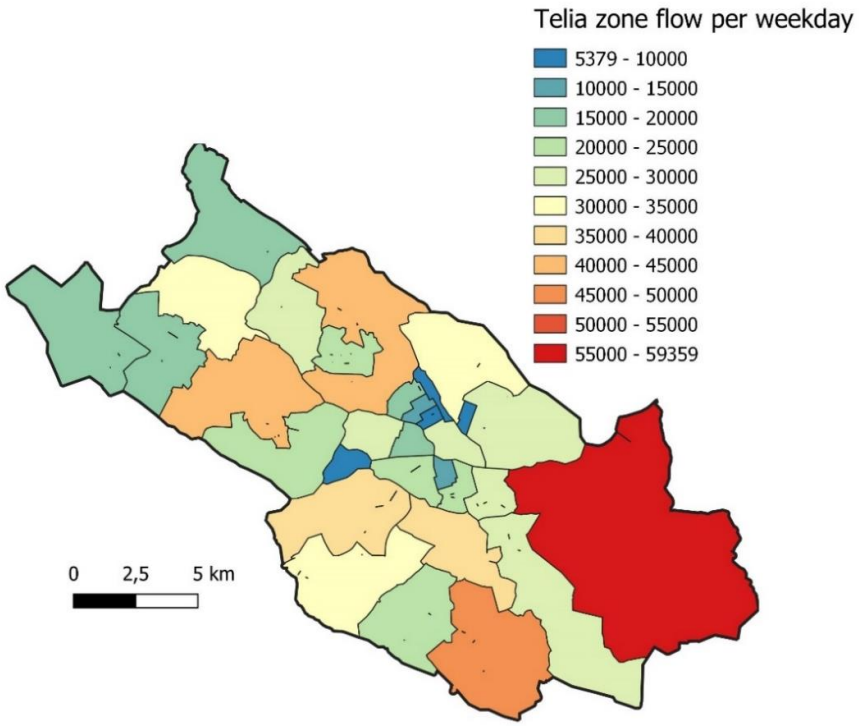


Figure 26 Spatial distribution of Telia zone flow.

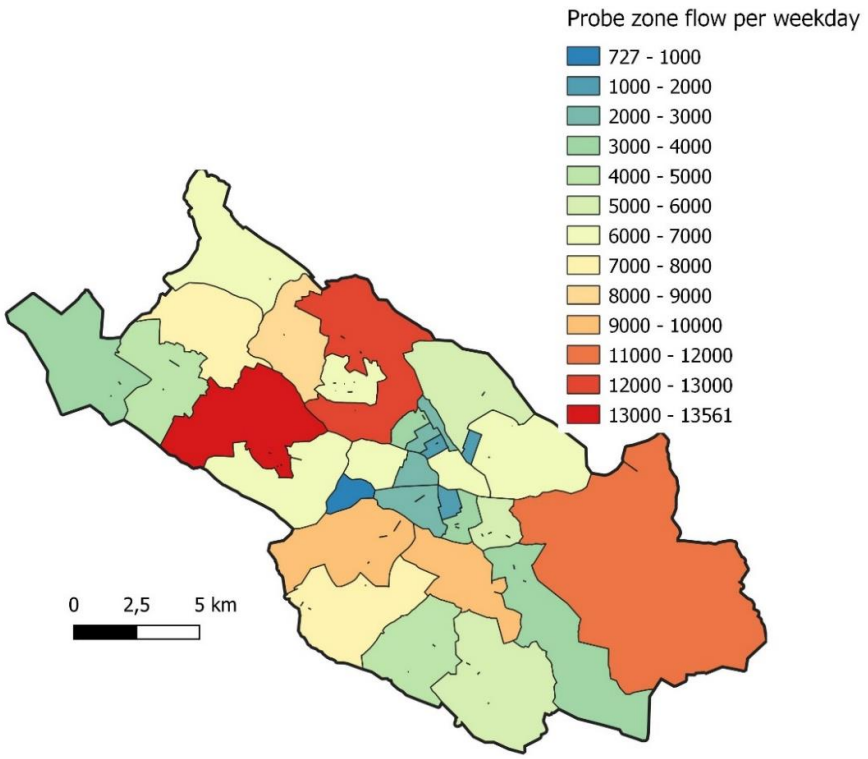


Figure 27 Spatial distribution of GPS probe zone flow

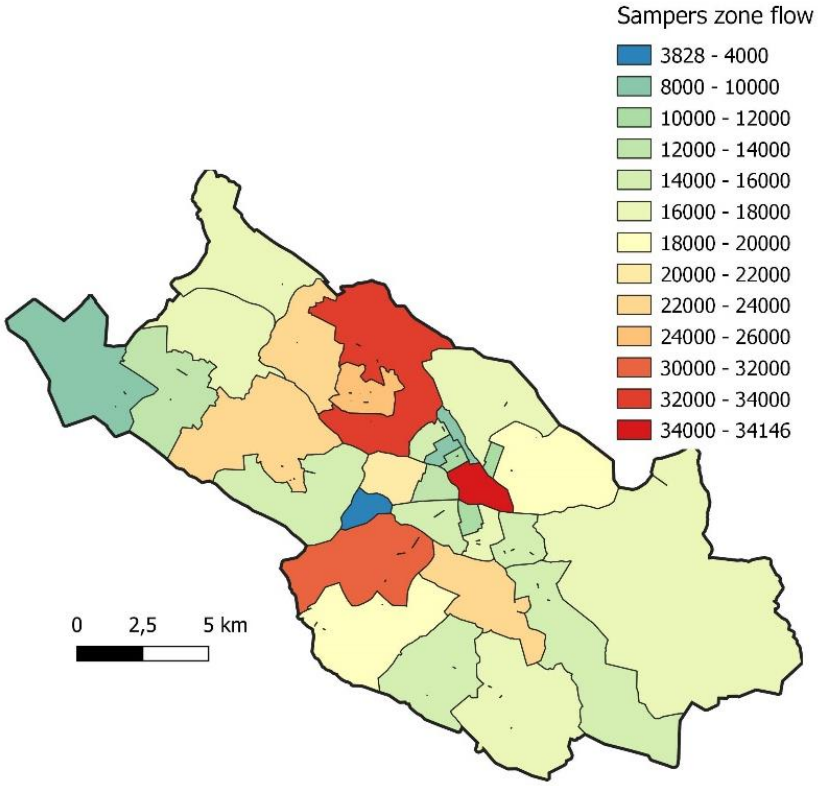


Figure 28 Spatial distribution of Sampers zone flow.

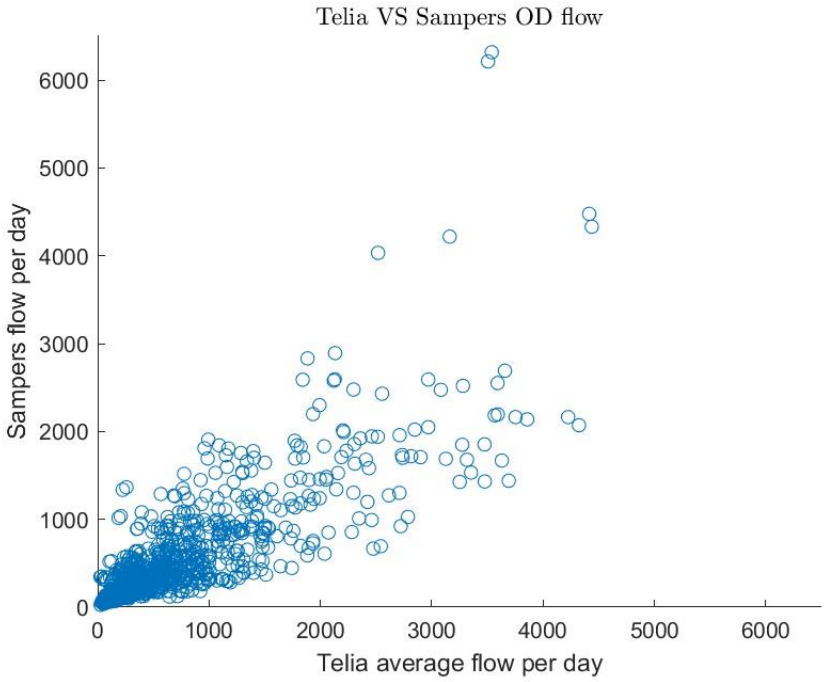


Figure 29 Comparison of average Telia OD flow for an average weekday and SAMPERS flow.

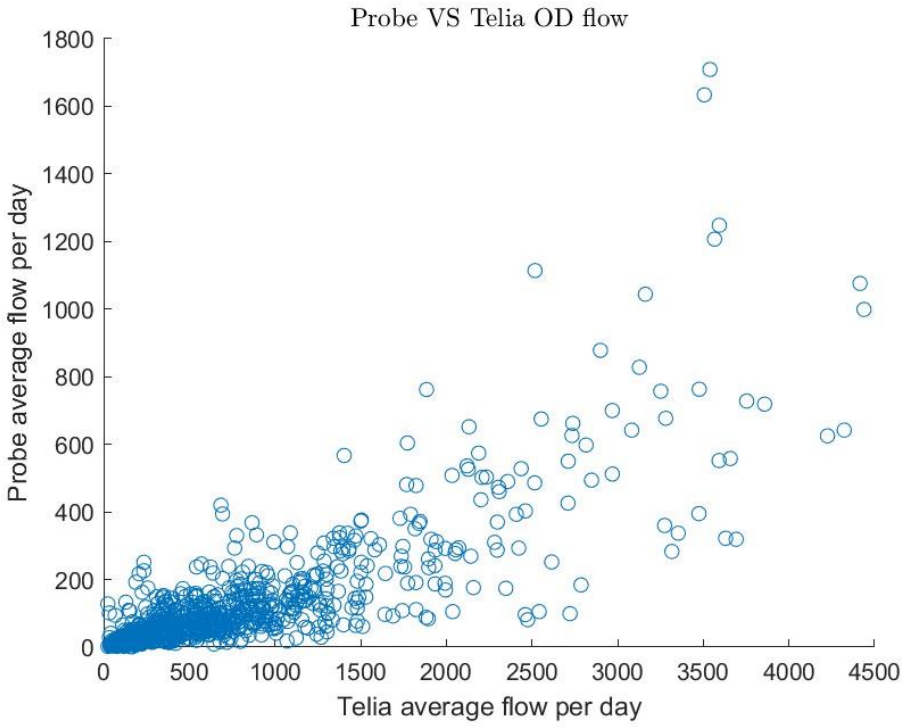


Figure 30 Comparison of GPS probe flow for weekdays during 5 weeks and Telia OD flow.

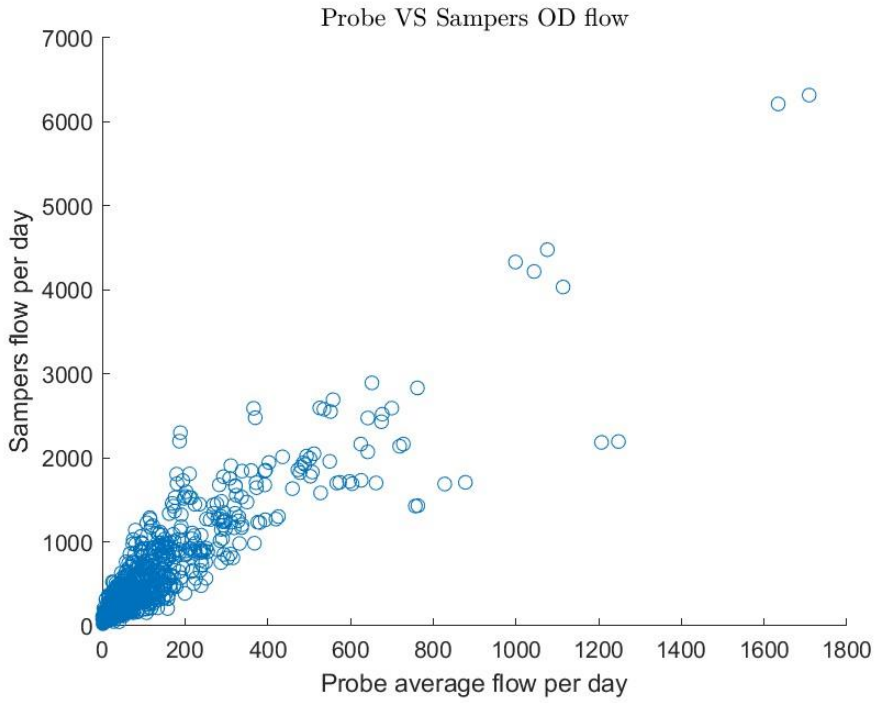


Figure 31 Comparison of GPS probe OD flow for weekdays during 5 weeks and SAMPERS flow.

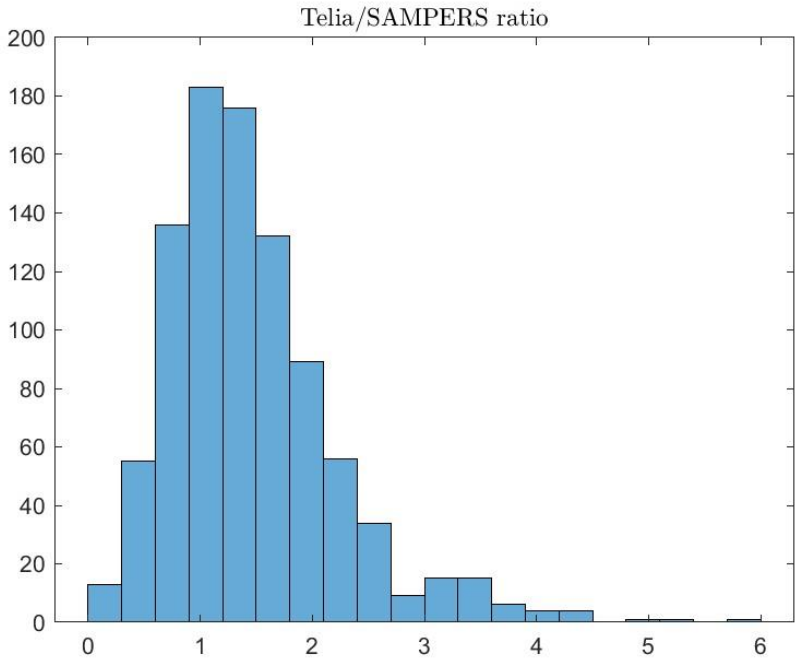


Figure 32 Histogram of Telia/SAMPERS OD flow ratio.

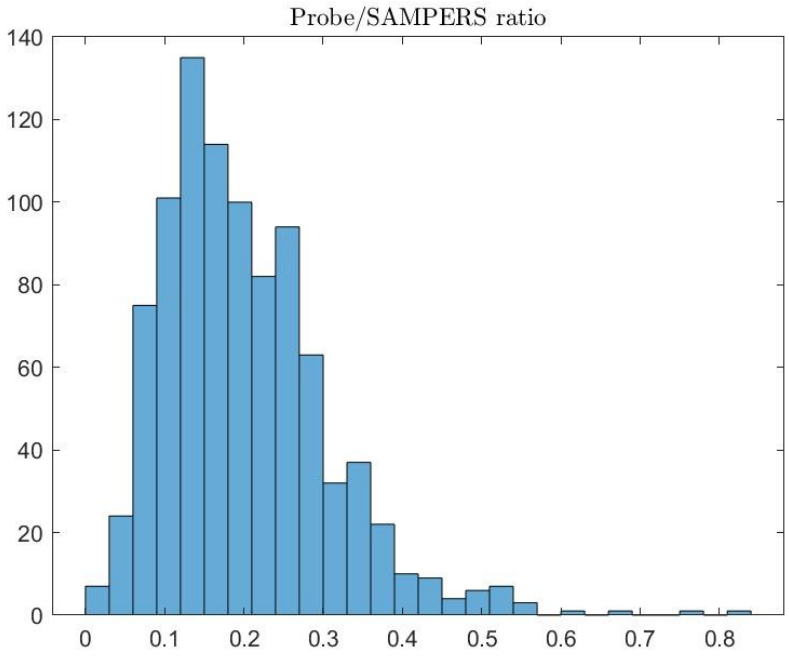


Figure 33 Histogram of Telia/SAMPERS OD flow ratio.

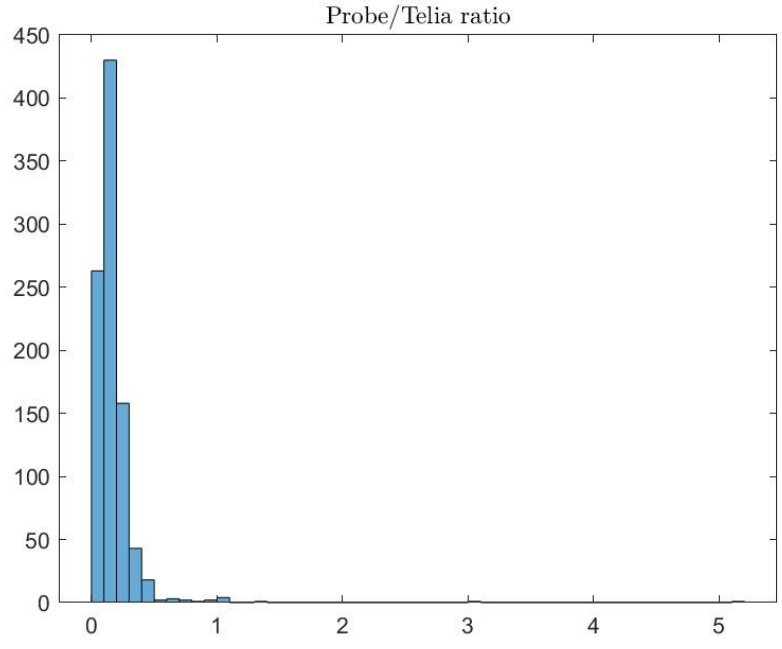


Figure 34 Histogram of GPS probe/Telia OD flow ratio.

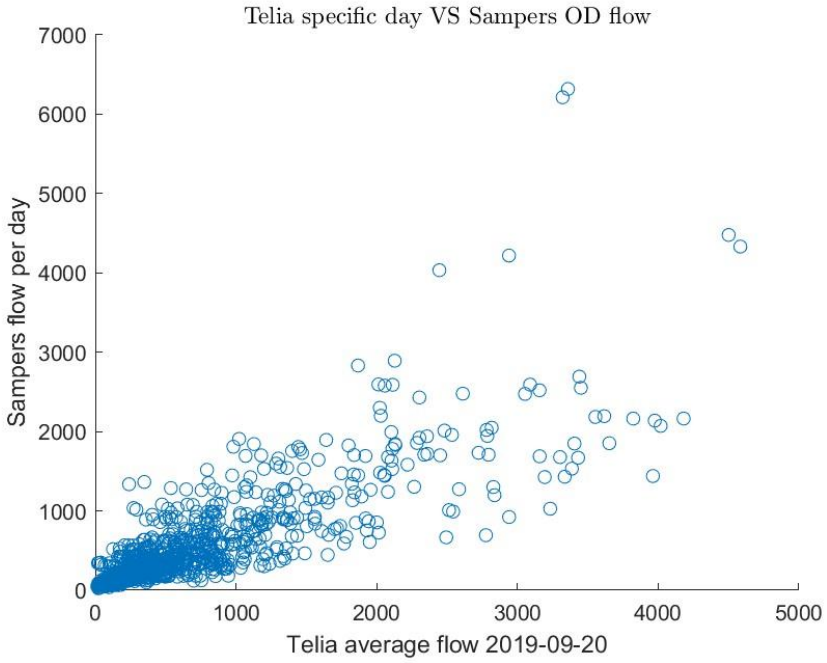


Figure 35 Comparison of Telia OD flow for a specific day and SAMPERS flow.

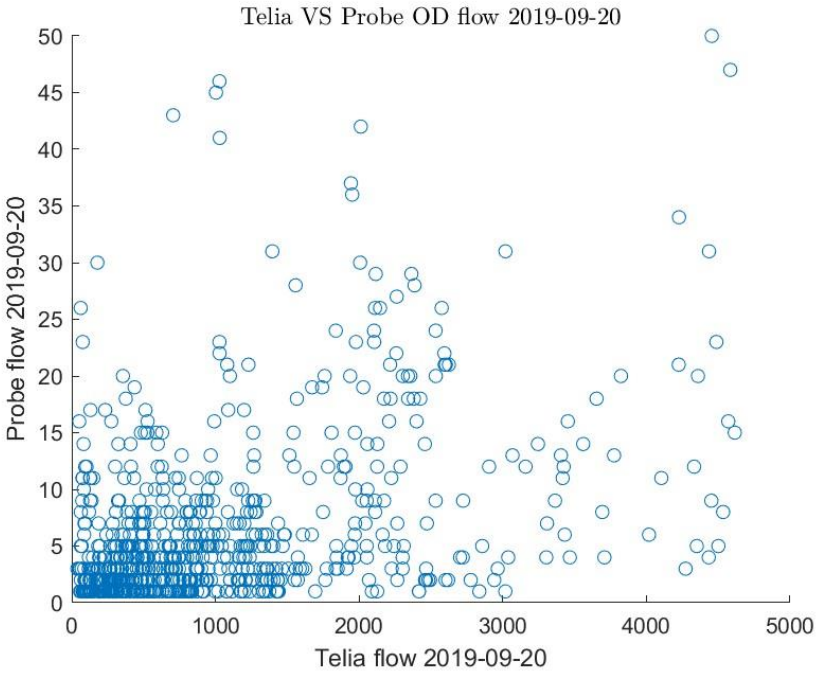


Figure 36 Comparison of Telia OD flow and GPS probe OD flow for a specific day.

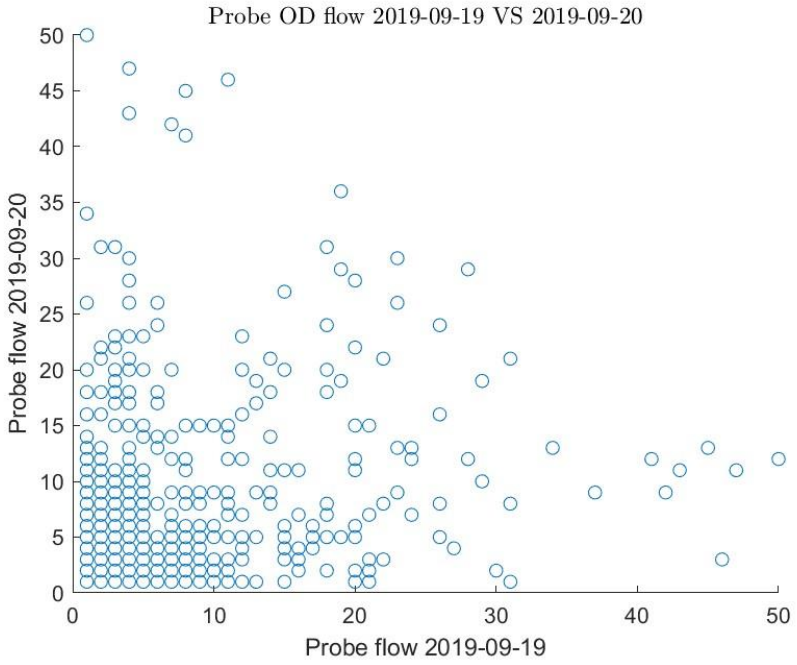


Figure 37 Comparison of GPS probe OD flow for two specific days.

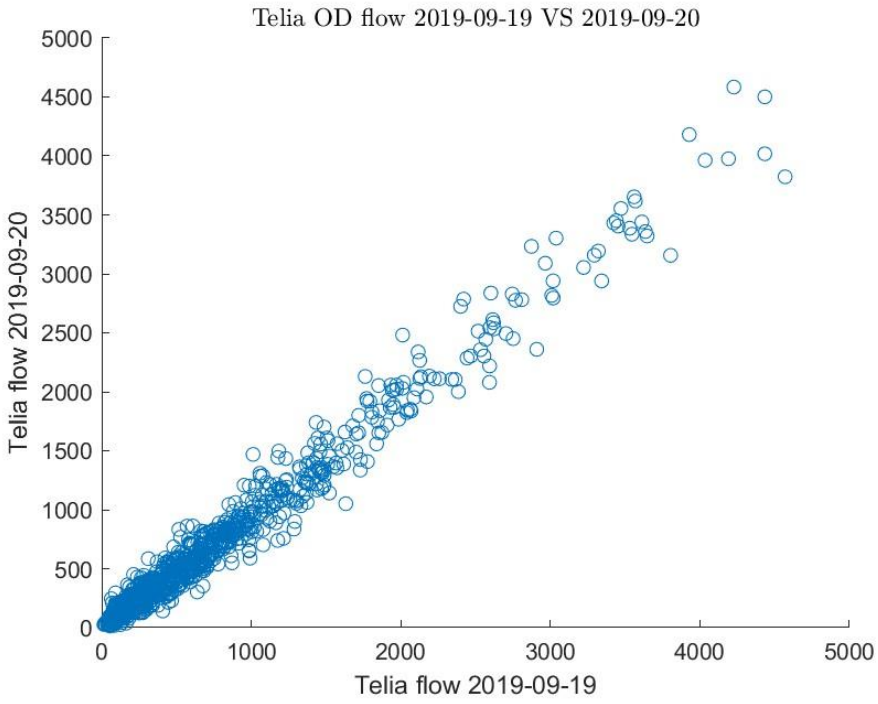


Figure 38 Comparison of Telia OD flow for a two specific days.

The results show that both mobile network data and GPS probe data have correlation to both the demand model SAMPERS and to each other for the larger Telia zoning. However, the GPS probe data needs aggregation over several weeks to get enough samples on OD level even for the larger

Telia zoning. The mobile network OD has a good correlation and reasonable variance between days, indicating that daily variations of mobile network data is promising and interesting to analyse further.

4.3 Major event detection

During the development of the methodology, three case studies of major events in Stockholm were investigated. Based on the specified areas of interest derived from the survey results and follow-up meetings, three events that occurred within the timeframe of available data, from September 16th to October 21st, 2019, were selected. These three events, namely Lidingöloppet, the Bicycle Carnival, and a climate demonstration, were chosen as Stockholm's case studies. It was considered likely that all of these events influenced route choices and potentially resulted in changes in Origin and Destination (OD) patterns, with a high probability that Lidingöloppet had an impact on OD-pairs. The selection of these cases was based on the clear and limited locations and timing, both in terms of time and space. Additional events were incorporated into the project as it progressed.

4.3.1 Analysis of case study areas

In each case study, a set of specific links was initially chosen for detailed examination in various aspects. This involved a thorough analysis of parameters such as speed, the volume of observations, and trip length. In the subsequent stage, the entire network in the given area was analysed for that particular date.

4.3.2 Network analysis of raw data

For the study of the selected case studies and major event detection and analysis, the dataset was used in its raw form and aggregated into 24-hour periods. The study area considered encompasses Stockholm, Linköping, and Norrköping, along with their surrounding regions. Starting from the base data of trajectories and trips, information for each trip was collected to establish connections between trip data and the corresponding links in the OpenStreetMap network (OSM). By associating the trajectories and trips with OSM links, it became possible to explore data such as flow rates for different links. Given the variation in penetration rates, various methods for measuring changes in the system were examined, including speed, distance traveled, and flow.

The primary approach for detecting changes involved comparing the traffic on links to that of an average day. Additionally, a combination of distance traveled (the total distance traveled in a straight line from point A to point B) and flow was used to identify variations compared to the average state. For each link and date, the following information was stored: the number of observed cars traveling on the road, the total distance traveled by the cars passing the link, the mean distance traveled, and the mean speed on the link.

The average values of these parameters were calculated for the entire period, establishing a "base" value to which each day's data could be compared. To enhance computational efficiency, links with a mean of fewer than 20 observations per day were removed. A link was flagged on a specific date if the deviation from the mean exceeded a threshold value. These links were then visualized using varying levels of saturation, with red indicating an increase and blue indicating a decrease, to illustrate the percentage of change compared to the average day. This method was used to investigate the three cases presented: Lidingöloppet, the bicycle carnival, and the climate demonstration.

4.3.3 Explorative approach

To identify new events based on the available data, an exploratory approach was implemented. The primary goal was to identify the most significant deviations while filtering out smaller, less significant changes. This method applied the same techniques used previously, involving a comprehensive search throughout the dataset for the most substantial deviations from the average state.

Through the application of this method, significant differences in the network across various areas were detected. Once these notable changes were identified, the focus shifted to connecting them with actual incidents or disturbances within the network. The top 100 links displaying the most significant recorded differences were specifically selected, and the gathered information was imported into the OpenStreetMap (OSM) network for visualization.

For further analysis, the continuous links that exhibited notable deviations from the average were further examined. These links were cross-referenced with dates and geographical locations mentioned in incident reports from Stockholm, as well as incidents reported online on corresponding dates.

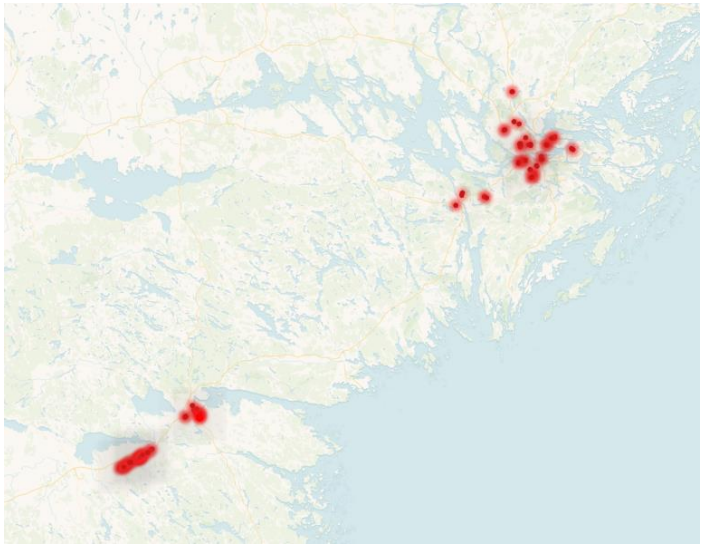


Figure 39 visualization of 100 links with the largest deviation from an average day.

One case was chosen for in-depth examination. This selection was based on the reported location of the accident and the identification of links that exhibited significant disparities. Two specific links were selected as a result. Subsequently, all vehicles passing through these links were extracted, along with details of all other links used by these vehicles. Using this dataset, an analysis was conducted to understand the route choices made by the vehicles passing through in response to the incident.

All links could be viewed to provide insights into how vehicles traversed them minute by minute. This approach enabled a detailed observation of minute-by-minute traffic patterns and route choices made by vehicles passing through the selected link at different time intervals.

4.3.4 Event study

Case 1 Lidingöloppet

Lidingöloppet is a weekend long meetup with over 20 runs and around 60,000 participants every year. It was chosen for study due to its considerable size and strategic location. The race is situated on an island close to the inner city of Stockholm. The location makes it an ideal event to study as we can expect additional trips to the island via one passage, the Lidingö bridge. The events time frame is well defined, with events from midday Friday the 27th until 17.00 Sunday 29th. The event has a long history and attracts large number of participants from all of Sweden and would be expected to influence the OD-matrix.

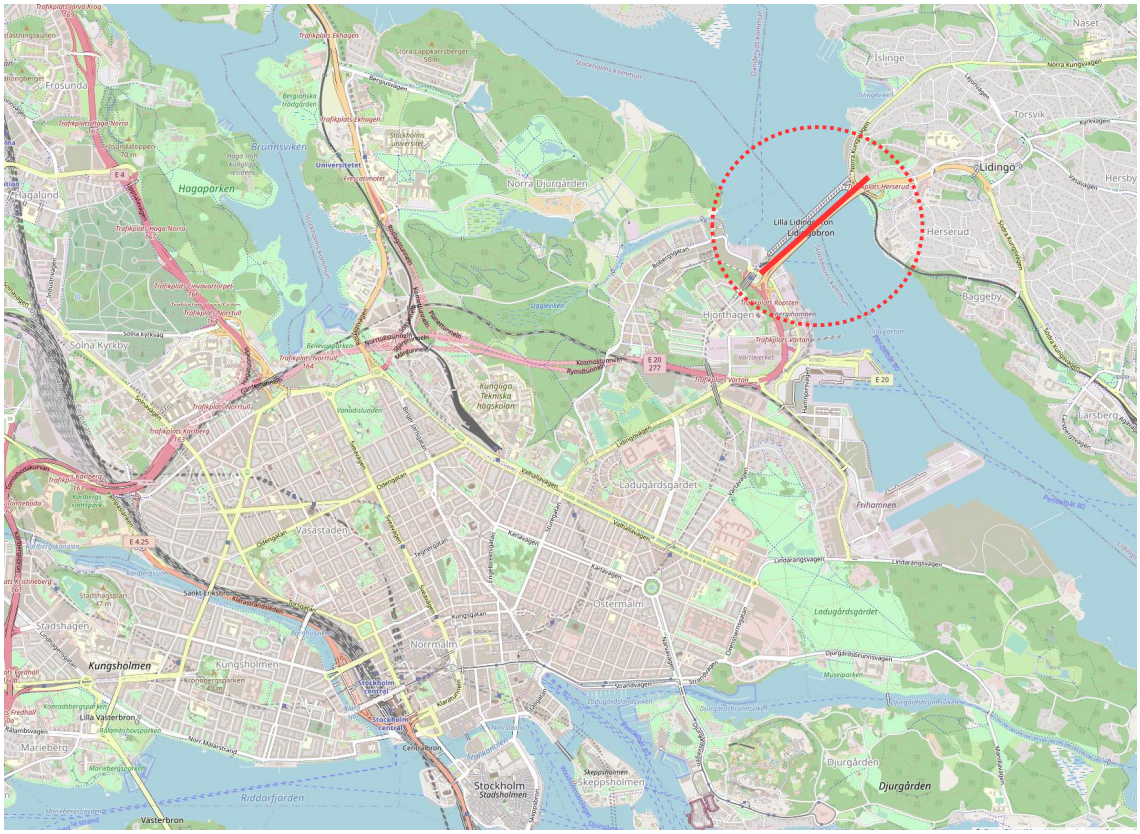


Figure 40 Lidingöbron, bridge that connects Lidingö municipality and the mainland.

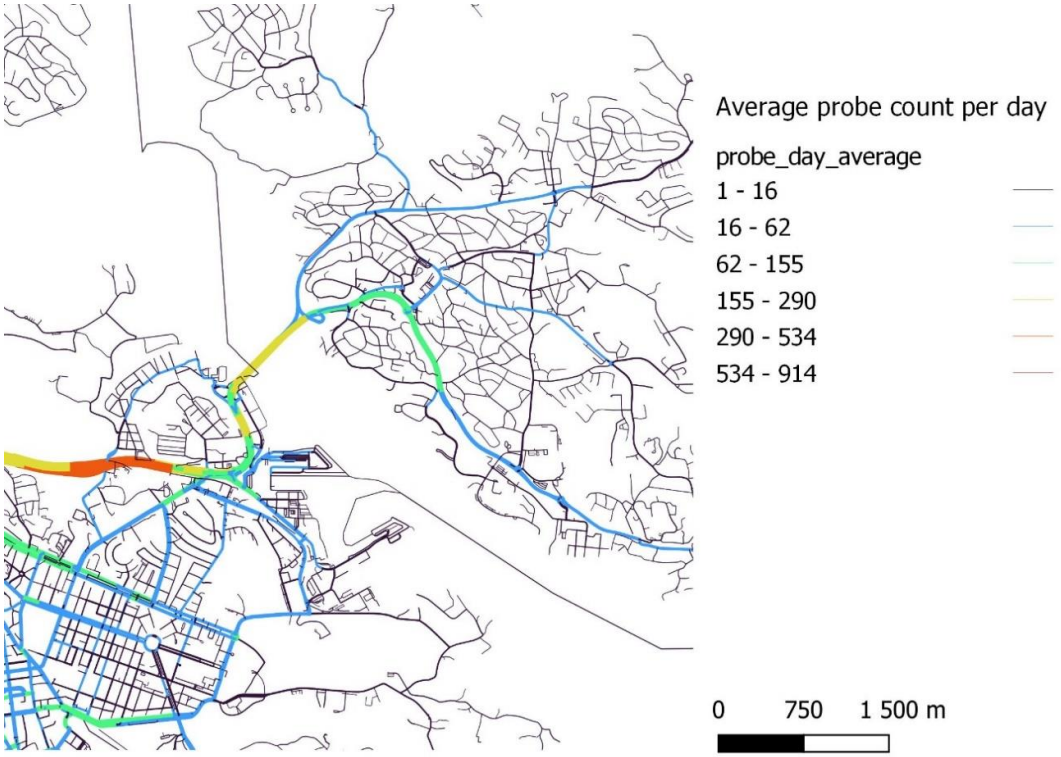


Figure 41 Average probe count per day.

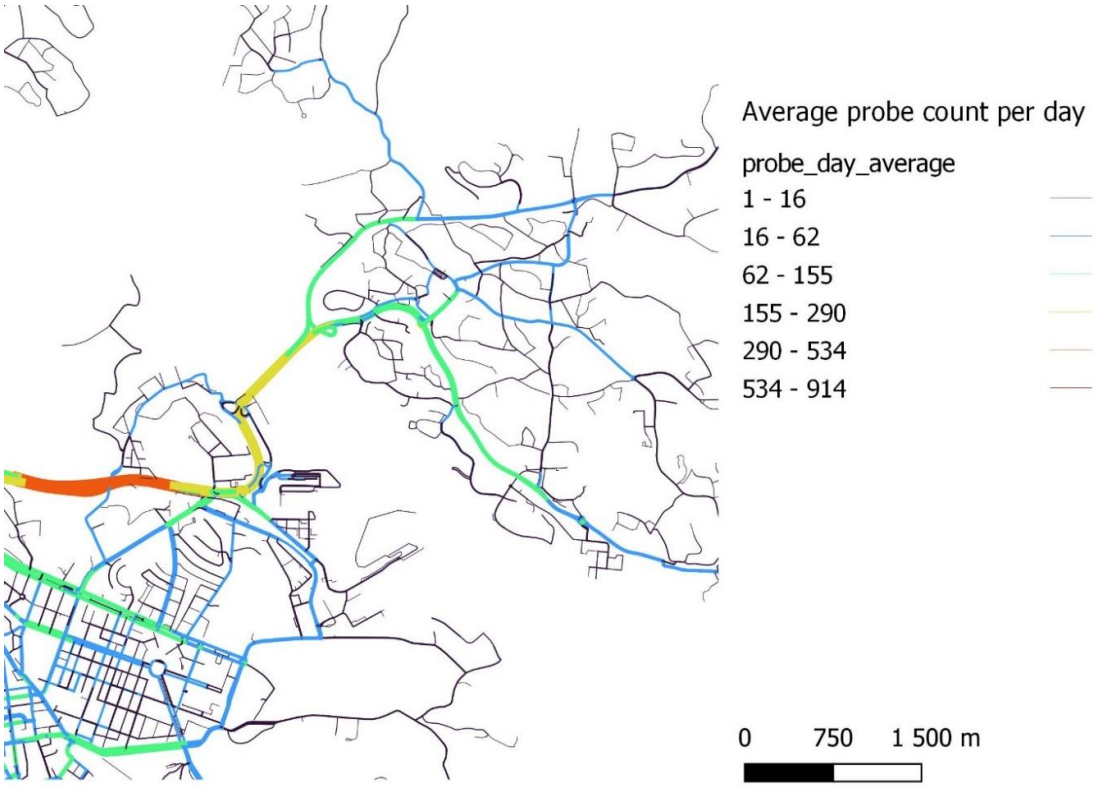


Figure 42 Probe count 2019-09-27 (Lidingöloppet Friday)

As the Lidingö bridge is the connector between Lidingö and Stockholm it was first chosen for a detailed study. Both directions were studied comparing the result for total distance traveled by cars using it during the 24-hour periods. The total distance traveled by the cars passing the Lidingö bridge during the time period is displayed in Figure 43 below. The date for the start of the event can easily be seen standing out, with a notable rise in comparison with other dates.

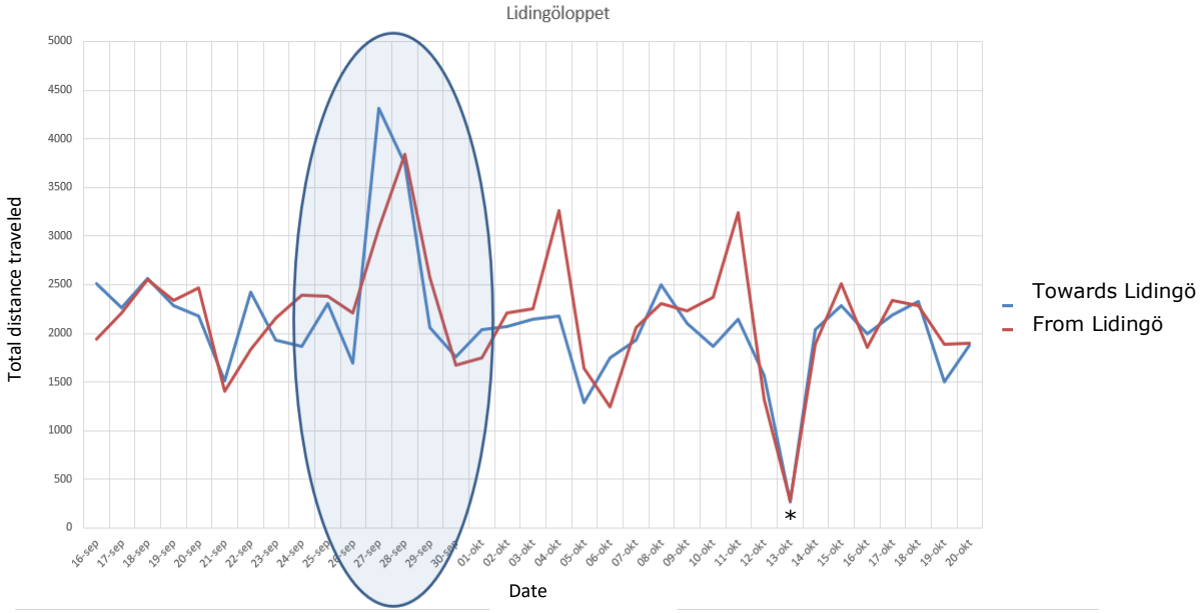


Figure 43 the total distance traveled by the cars passing the Lidingöbridge between 16th of september and 20th october.

*The low values for the 13th of october has been identified as a period with loss of data, some hours are not included for the 13th of october in the original raw data.

In addition to the link towards and from Lidingö the whole network was explored for the 27th of September. This resulted in Figure 44 depicting links in Stockholm and close municipalities flagged as relevant in the analysis. In the upper right corner we see Lidingö and Lidingö bridge with between 50 and 500 percent change. Furthermore, the large event and movement on Lidingö was also found with the explorative method presented.

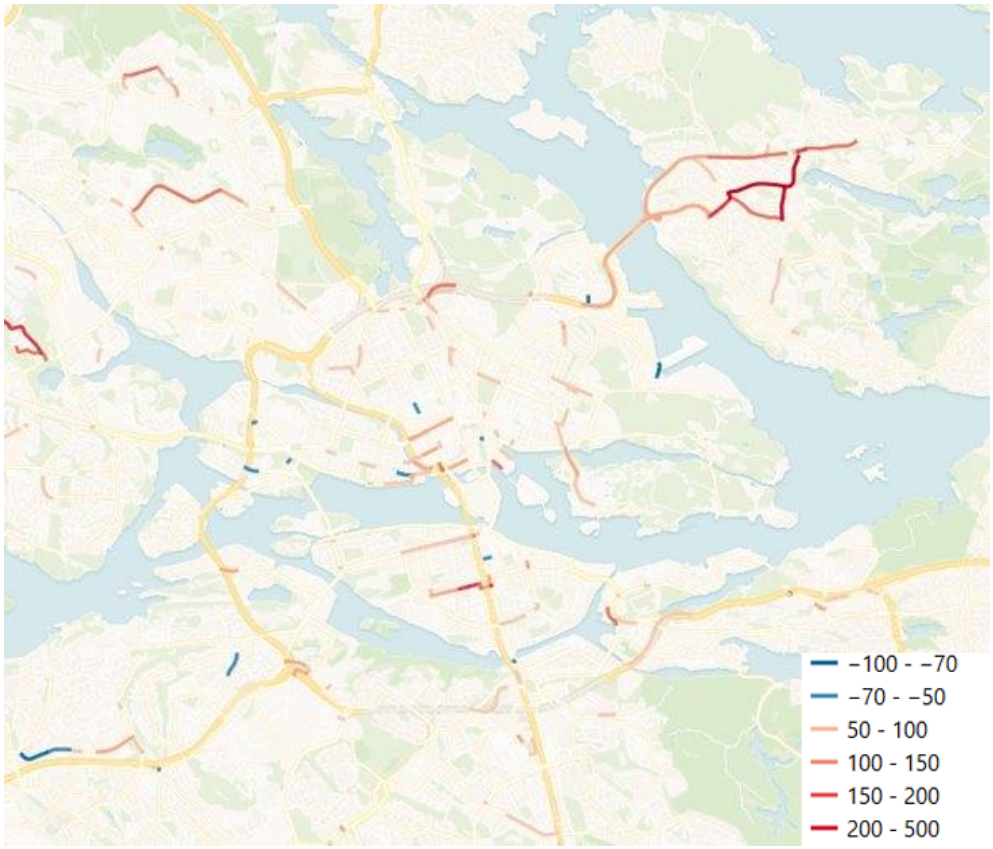


Figure 44 The difference in total travel distance in Stockholm the 27th of September 2019 compared to an average day.

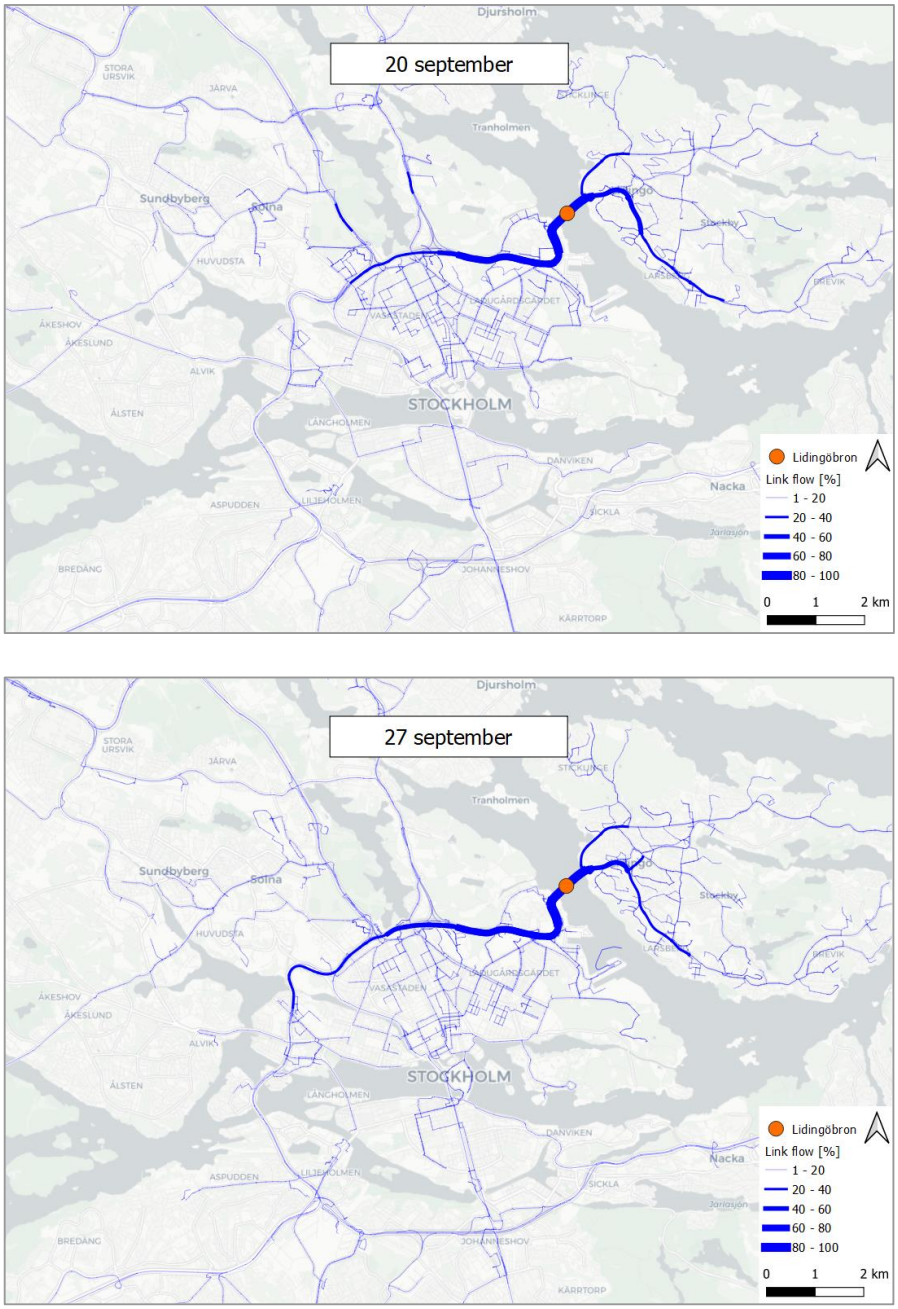


Figure 45 Route shares during Lidingöloppet Friday and the week before

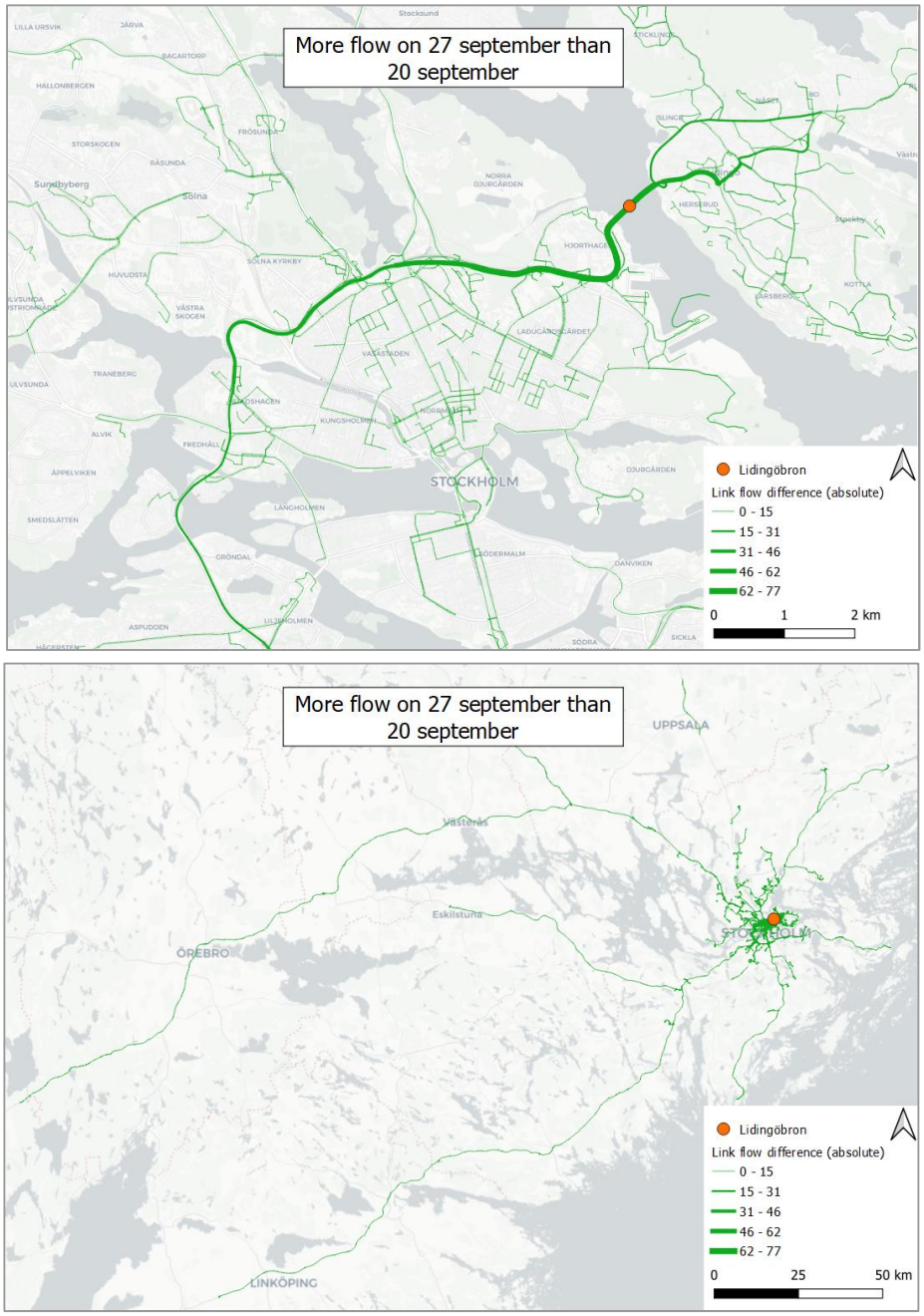


Figure 46 Links with more flow that passes Lidingöbron towards Lidingö on September 27 compared to September 20.

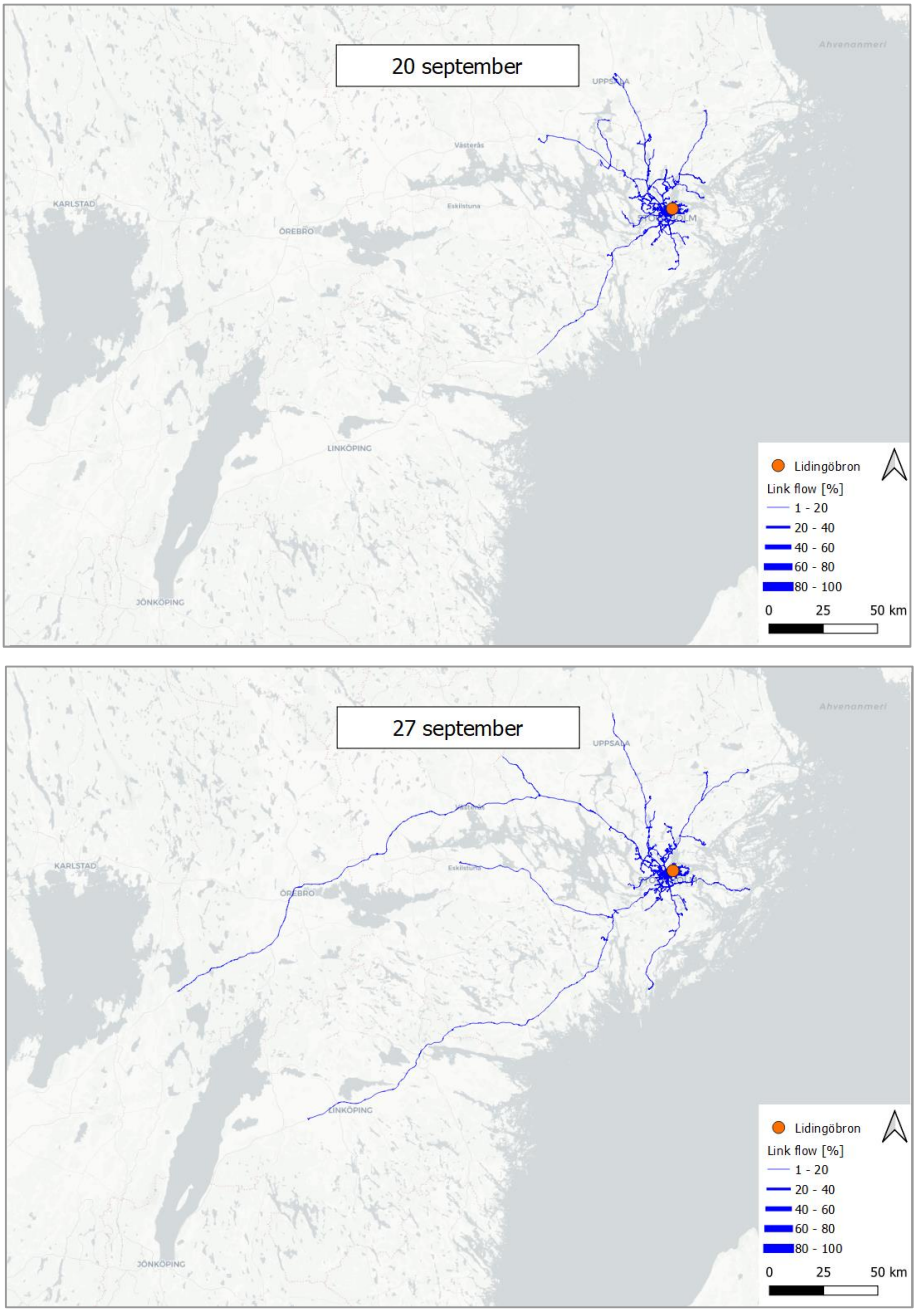


Figure 47 Route shares for longer trips that passes by the Lidingö bridge during 2019-09-20 and 2019-09-27, respectively. A larger share of longer trips can be observed for the Lidingö race weekend.

Case 2 The Bicycle Carnival

The second case is the bicycle carnival in the inner city. The 22nd of September 2019 as a part of the European Mobility Week a bicycle carnival was performed in the inner city of Stockholm between 14.00 and around 15.30 from Norrbron via City, Kungsholmen and Vasastan and back to the start again. Norrbron is one of the connections between the old town and the north of the inner city was closed off between 10.00 and 18.00.

Figure 48 Route of the bicycle carnival in central Stockholm.

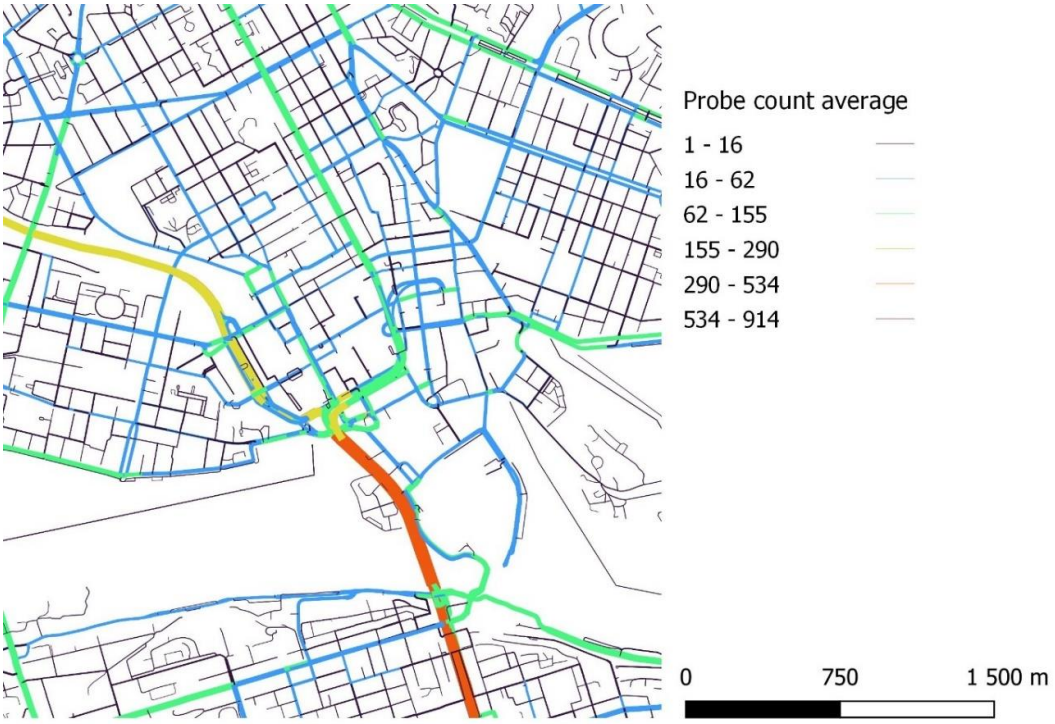
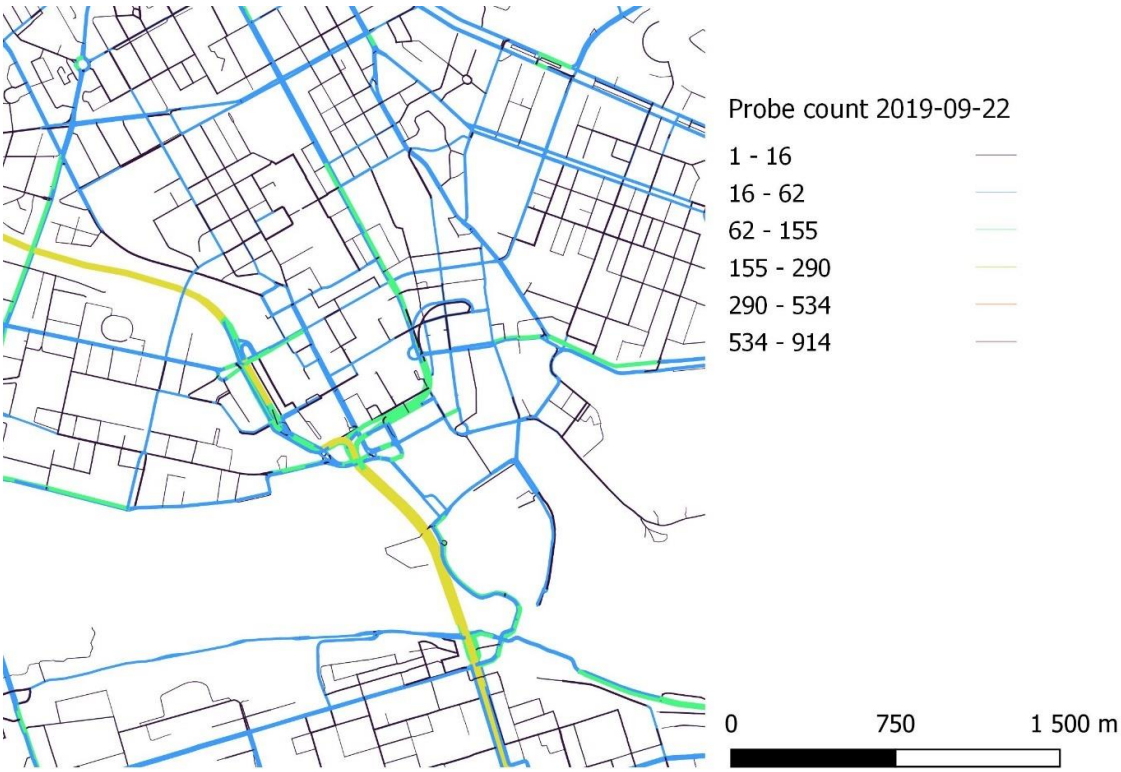


Figure 49 Average probe count per day



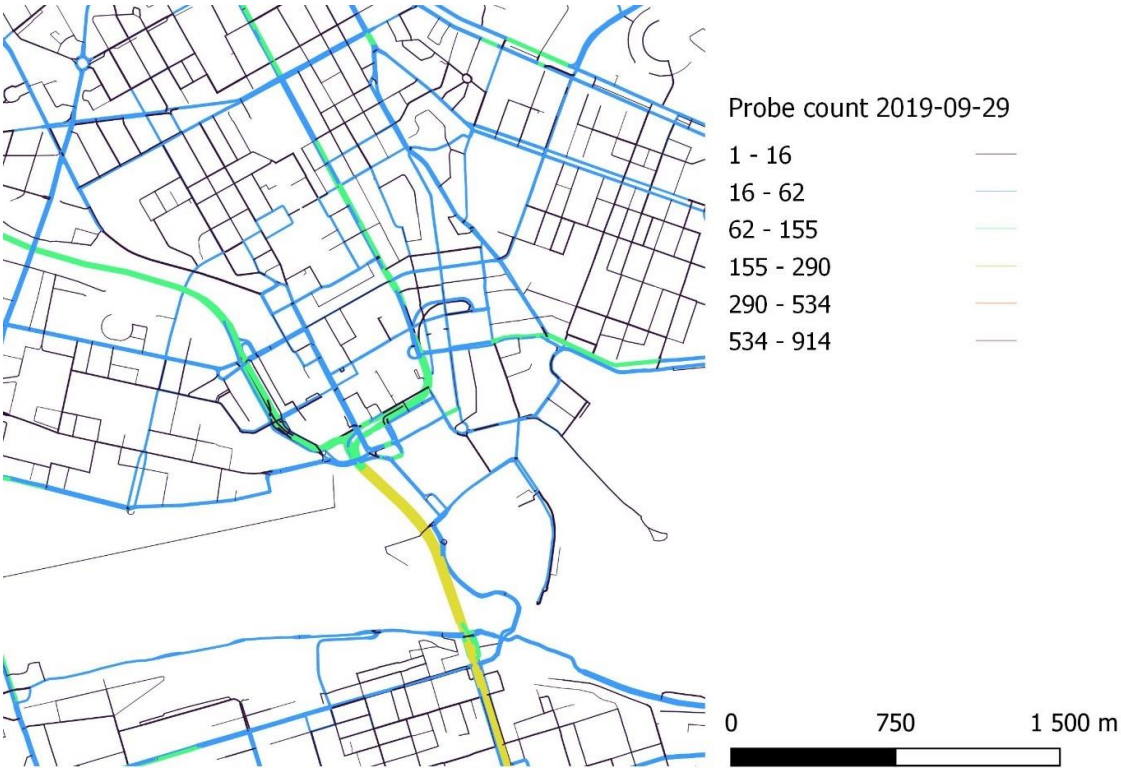


Figure 50 Probe count September 22 and 29.

Case 3 The Environmental Demonstration

On Friday, the 20th of September 2019, a demonstration for the environment took place in the inner city. The demonstration was assumed to include approximately 5000-6000 youths. Two different start times were provided: 13.20 and 14.00. As a result of the demonstration, public transport, including buses, was limited in the area, and access was restricted between 13:00 and 15:00. This event is presented as affecting public transport (buses). The demonstration route in the old town is unavailable. Three different alternatives are presented in Figure 28 below.

Figure 51 Route for the demonstration for the environment.

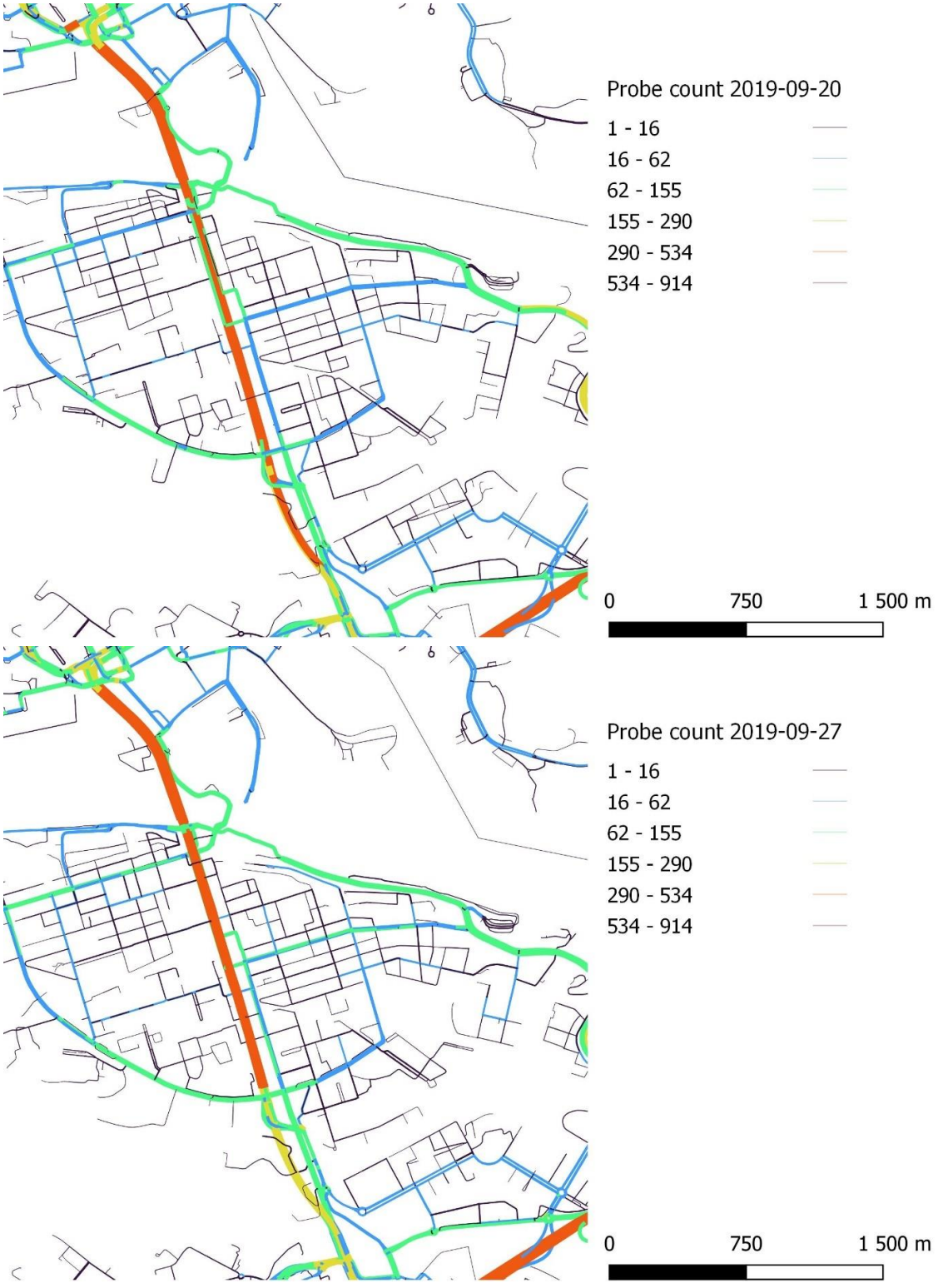


Figure 52 Probe count September 20 and 27.

Detected events

After cross-referencing some of the areas with dates and geographical locations from incident reports in Stockholm, as well as incidents reported online on the corresponding dates, three new cases were identified. These cases included an accident along with a closed tunnel during the morning rush (Case 4), an incident involving debris on the road in Stockholm (Case 5), and a car accident between a personal vehicle and a truck on E4 between Linköping and Norrköping, near Rystad (Case 6).

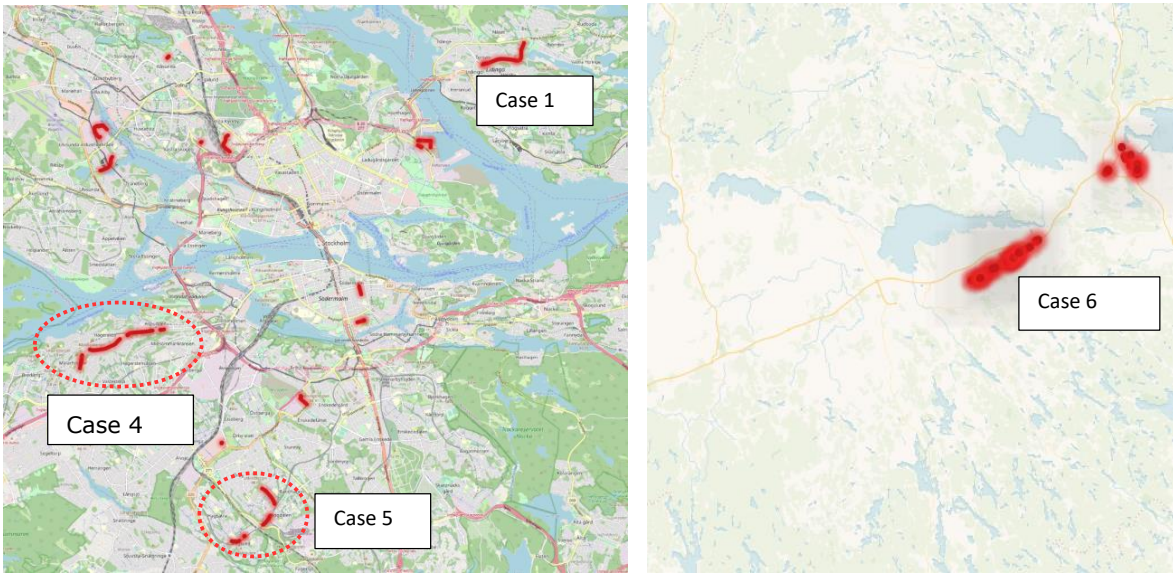


Figure 53 visualization of links in red with the largest difference in Stockholm, and the new case close to Rystad.

The visualization of the 100 links with the highest percentage change in total distance travelled also connected back to the first case (Lidingöbron) as seen in the upper right corner in figure. On the 27th of September, some links to Lidingö also appeared, the same date as the start of the event.

Rystad

The area around the old E4 close to Rystad was selected for further study. Around 14.00 the rescue services were notified of a traffic accident on E4 around Rystad. A truck drove into a passenger car and the railing. While the rescue services worked the road was closed in the north bound direction and there was limited flow in the south bound direction. The accident caused long drawn-out queues. Following the accident, the police diverted traffic via the old E4 road between the roundabout Mörtlösa and Norsholm. The diversion of traffic and increase in trips across the old E4 could be what the explorative process picked up, as increased amounts of travelers who traveled longer trips used the parallel road.

Figure 54 Identification of major event between Norrköping and Linköping.

The event and the noted links are displayed below. Cars that pass the link on E4 during their route are displayed in yellow, and cars using roads where cars from the E4 were rerouted to are displayed in green.

In the first illustration of route choices, we see the area and routes during the 60 minutes before the accident was reported. After the accident was reported, the routes changed, with more cars covering longer distances between their origins and destinations by using the old E4. Cars starting on the E4 link could also indicate queues, as stopped vehicles in the dataset can become new trips, suggesting that these might be cars stuck in traffic. The third illustration of routes used shows the 45-minute period when no cars in the dataset passed the E4 link after the reported accident.

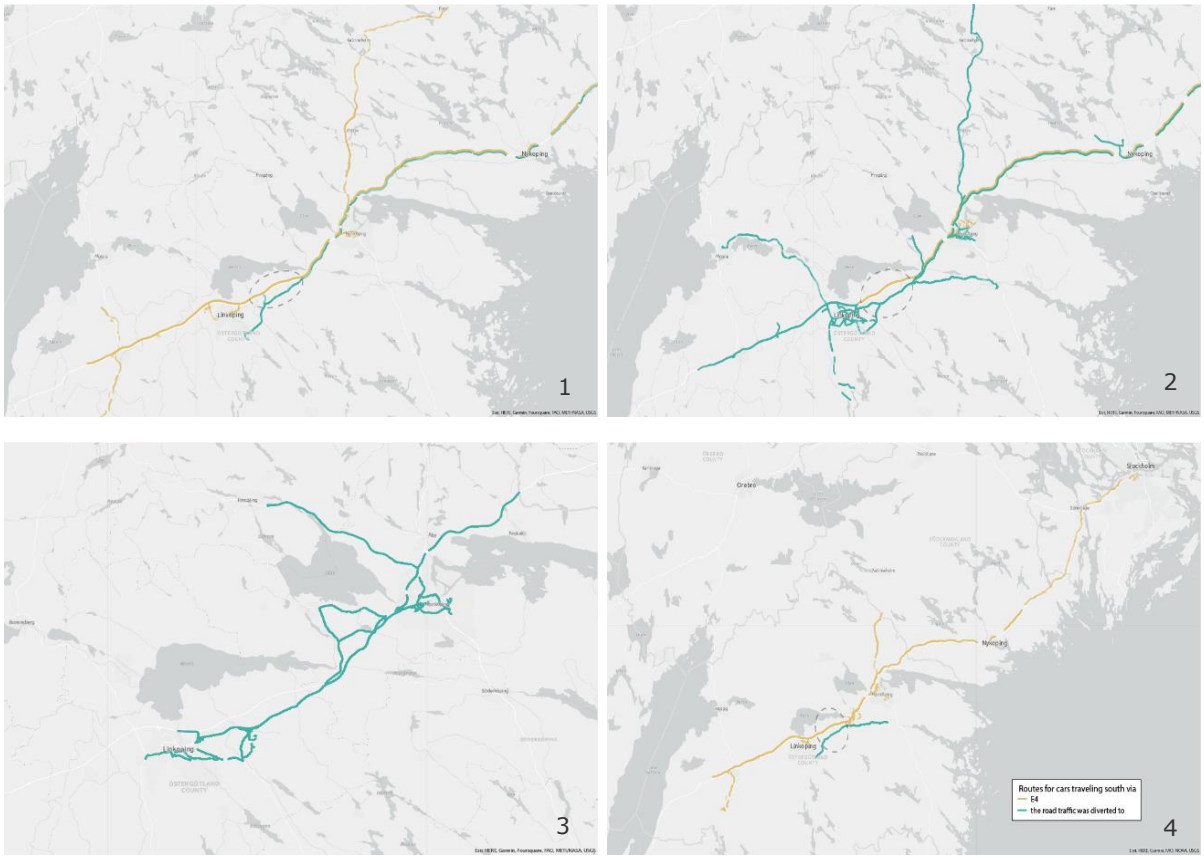


Figure 55 the routes used before during and after the time of the accident.

4.4 Use cases in Nyköping and Gothenburg

The study of Nyköping primarily focused on descriptive statistics and analyses, with the intent of contrasting it with Stockholm as a smaller city. The analyses in Nyköping centered around parameters such as flow/number of observations, Origin and Destinations, and flow distributions between two interchanges. The study in Gothenburg, on the other hand, emphasized freight analyses, in addition to the whole vehicle fleet.

Nyköping

The examination of Nyköping prioritized the application of Inrix data for descriptive analysis within the context of a smaller city. The Inrix dataset used for the Nyköping case study corresponds with the same dataset as the rest of the study, spanning from September 16, 2019, to October 20, 2019.The data from Inrix was initially combined with classifications from OpenStreetMap, such as maximum speed allowed and road type (primary, secondary/tertiary). The number of observations was aggregated into 24-hour flows, facilitating the creation of an interactive tool that presents this information for easy visualization of the Inrix data."

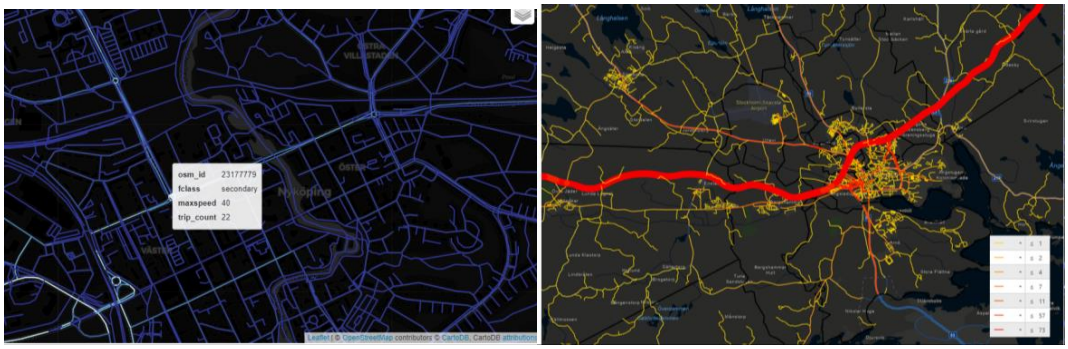


Figure 56 Visualization and interactive map of classification on road network and flow.

For the analysis of OD pairs, Inrix waypoints and their sequences were utilized to generate OD-desire lines. The zones were reconfigured for the major roadways leading out of the city boundary. The desire lines, as depicted in Figure 57, illustrate the movement of tripIDs through these zones

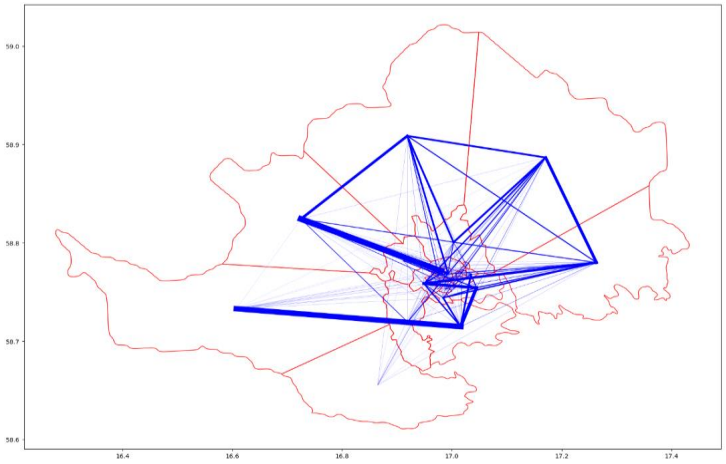


Figure 57 OD desire lines in Nyköping.

For a descriptive analysis of route choice and flow distributions, the aggregated flows between interchanges A (Trafikplats Hållet) and B (Trafikplats Björshult) were examined. Maps displaying trip counts for the considered road segments in Nyköping city were initially created, with variations in color and thickness.

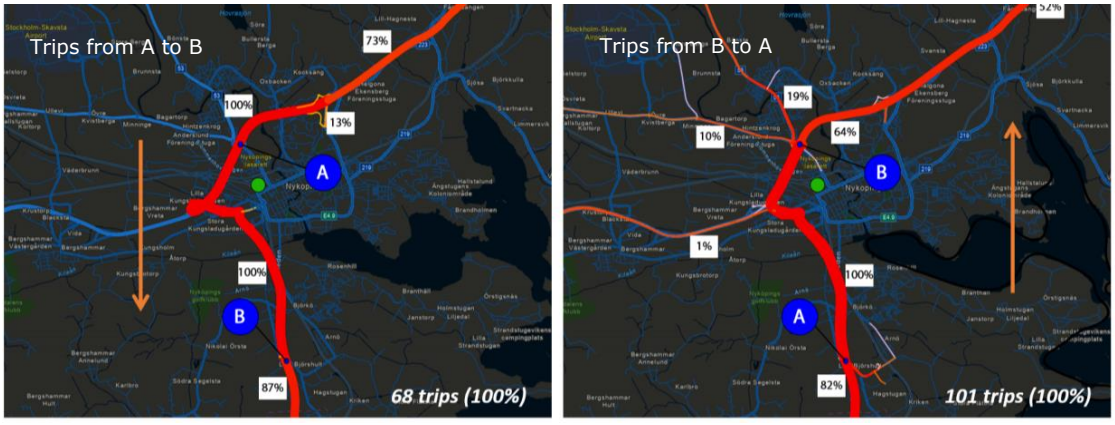


Figure 58 Aggregated monthly flows south bound and north bound (16 Sep – 16 Oct).

From September 16th to October 16th, the modal shares were distributed as follows: personal cars accounted for 54%, light trucks for 45.5%, and distribution trucks for 0.5%. The percentage of trips that occurred on an average day was higher from Monday to Thursday compared to Friday and the weekend.

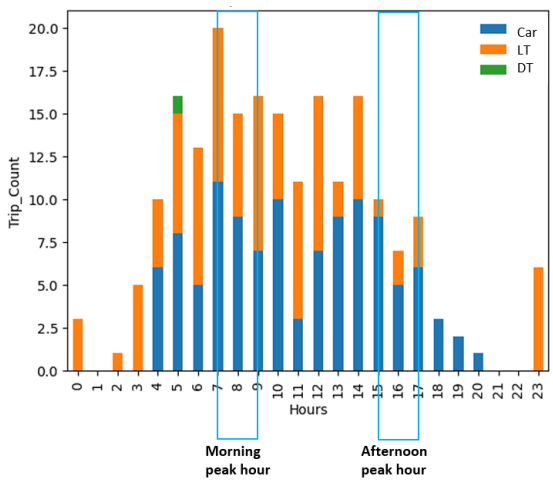


Figure 59 Trip count for each hour of an average day.

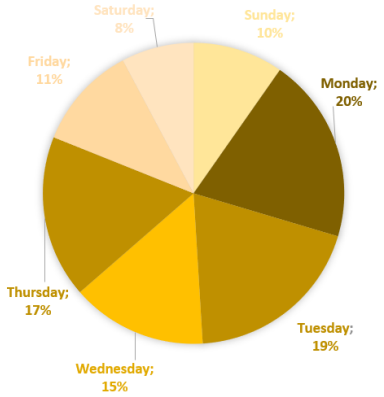


Figure 60 Distribution of trips over the days of the week.

The conclusions drawn from the division between days of the week, modes, and routes, which were aggregated over the entire month of data, can also be applied to the visualization of trips between points A and B during morning and afternoon peak hours. These trips were visualized in the same manner as those occurring throughout the entire day. However, the division of aggregated days into three-hour slots revealed that the dataset around the city of Nyköping is too limited to be subdivided into such time slots. In this case, a three-hour slot only yields 7-21 trips. Consequently, the penetration rate for the 2019 dataset is insufficient for this level of analysis.

Freight transport analysis Gothenburg

With access to INRIX GPS-probe data in Gothenburg, the project partly aimed to analyse route choices for various vehicle classes, primarily classified based on the vehicle's weight. Using this data, travel distances distributions was extracted for both passenger car traffic and heavy traffic, including freight transportation. This information provided the unique insight into how different vehicles interacted with the city's road network.

In the analysis, a significant change in two different datasets was observed, representing different time periods. A pivotal event that significantly impacted the infrastructure was the opening of the Mariheholm tunnel, which created entirely new conditions for the later time period. See Figure 61 were the red circle highlights this area. This had a direct impact on route choices and traffic patterns in the city.

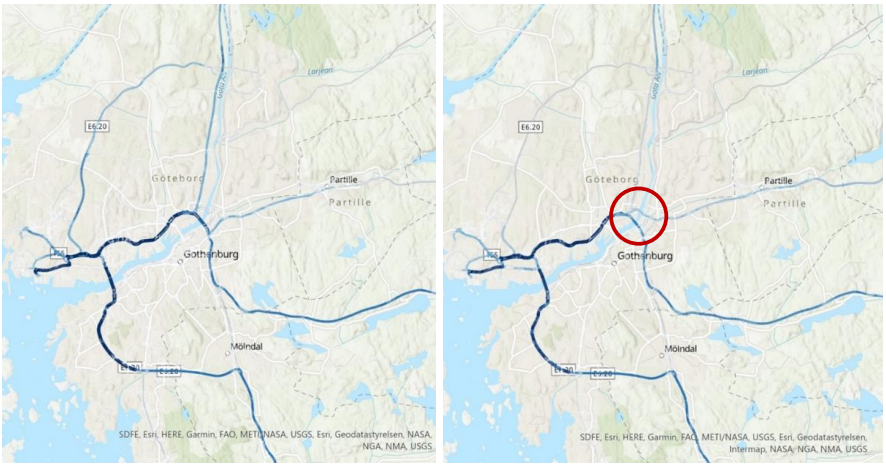


Figure 61 Route choices for heavier vehicles during the given time period.

By being able to track and analyze these types of changes, we can better understand and predict the dynamic needs and challenges of the city's traffic system. The benefit of being able to classify vehicle types and their relationship to the current state of the road network can be further multifaceted, making it possible to tailor and optimize the road network for different types of vehicles, especially heavy traffic, and freight transportation. By considering the specific needs and limitations of different vehicle types, traffic flow could be improved, road wear reduced, and traffic safety enhanced. Further insights were offered by conducting a "select link" analysis, where specific sections or road areas were closely examined in relation to their connection to various destinations for vehicles and freight transportation.

This material can, for example, be further used to investigate the road network's load-bearing classification. By observing how heavier vehicles use the road network based on weight classes, a comprehensive overview of the need for road classification can be provided. Usually, this knowledge is unknown, that is, which weight class travels where in the road network, and one needs to engage in discussions with various organizations to find this out. Alternatively, conduct extensive measurements, which is a qualified guessing game regarding which road segments to measure when you don't know in advance which roads heavier vehicles use that should not be used. With vehicle data, route choices for different weight classes are mapped across the entire road network simultaneously. The data provides a unique insight into which roads have insufficient standards based on the traffic traveling on them. It offers the city and decision-makers a valuable opportunity to gain a comprehensive overview of the road network, where the question of increased load-bearing classification or new policies for freight transport can be introduced.

5. JOINT OD AND LINK FLOW ESTIMATION

Time-dependent Origin-Destination (OD) matrices are an essential input to transportation models. A cost-efficient and widely used approach for estimating OD matrices involves the exploitation of flow counts from stationary traffic detectors. This estimation approach is also referred to as assignment-based OD matrix estimation because, typically, Dynamic Traffic Assignment (DTA) models are used to map the OD matrix to the link flows. The conventional DTA establish a complex non-linear relationship between the demand, and the link flows, adding an inherent complexity to the OD matrix estimation problem. In this section, attempting to exploit the growing availability of GPS probe data, we suggest a solution approach that is based on a Data-Driven Network Assignment (DDNA) mechanism. The DDNA utilizes GPS probe data to capture congestion effects, providing a linear mapping of the OD matrix to the link flow observations.

In [9] a more detailed description of the method is available as well as results of synthetic-data experiments. The remainder of this chapter presents initial results for the method in Stockholm.

5.1 DDNA

In DDNA a linear mapping between OD and link flows is estimated using GPS probe data. This linear mapping is an assignment matrix (***A***) that includes both route choice and route flow propagation (network loading). Detailed GPS probe data is used to estimate a route set and route choice matrix that estimates the proportions of OD flow that is using the different routes in an OD pair. GPS travel time data is then used as input to estimate a dynamic loading matrix that determines how much of a route flow that is propagated to a given link in a given time period. With the OD matrix ***X,*** the assignment matrix ***A*** *and link flows* ***Y,*** *there is a relationship between the link flows and the OD matrix according to* ***Y = AX.***

5.2 OD Estimation

The OD estimation problem based on DDNA utilizes a linear mapping (***A***) between OD and link flows to connect OD flows to link flows. However, the OD estimation problem based on link counts is an underdetermined problem with many potential OD matrices that can generate the observed link flows. To reduce the problem of having many alternative OD flows a target matrix can be introduced that is reflecting the structure of the OD matrix. With a target matrix and link flow observations, the OD estimation problem based on DDNA can be formulated as

$$\boldsymbol{X}^* = \operatorname{argmin}_X \left[\left\| \widehat{\boldsymbol{X}} - \boldsymbol{X} \right\|_2^2 + \left\| \widehat{\boldsymbol{Y}} - \boldsymbol{A}\boldsymbol{X} \right\|_2^2 \right]$$
$$\text{s.t.} \quad \boldsymbol{X} \geq \boldsymbol{0}$$

where

$\boldsymbol{X}^*$: OD demand

$\widehat{\boldsymbol{Y}}$: Observed link flows

$\boldsymbol{Y}$: Modelled link flows, $\boldsymbol{Y} = \mathbf{G}(\mathbf{X}) = \mathbf{A}\mathbf{X}$

$\widehat{\boldsymbol{X}_{\boldsymbol{1},\ldots,\boldsymbol{n}}}$: Target Demand

Figure 62 gives an overview of the OD estimation process where the OD matrix is adjusted so that the predicted link flows match the observed link flows as good as possible.

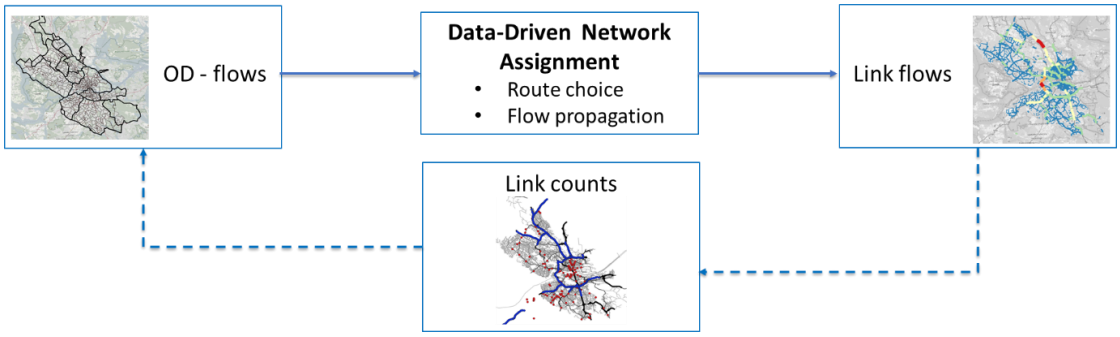


Figure 62 Overview of OD estimation based on DDNA

5.3 Data fusion

OD matrix estimation is a very challenging problem, and it is important to combine different data sources efficiently. In the previous formulation we combined the link flow observations with a target matrix from a demand model (SAMPERS). However, it is also possible to include other types of data in the optimization problem, for example mobile network data. We can combine multiple data sources by incorporating them in the objective function of the optimization problem

$$\boldsymbol{x}^* = \text{argmin}_x [w\|\hat{\boldsymbol{y}} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + w_1\|\widehat{\boldsymbol{x}_1} - \boldsymbol{x}\|_2^2 + w_2\|\widehat{\boldsymbol{x}_2} - \boldsymbol{x}\|_2^2 + \cdots w_n\|\widehat{\boldsymbol{x}_n} - \boldsymbol{x}\|_2^2]$$
$$\text{s.t. } \boldsymbol{x} \geq \boldsymbol{0}$$

Where $\widehat{\boldsymbol{X}_{1,\ldots,n}}$ corresponds to for example demand model OD, mobile network OD and GPS probe OD observations.

5.4 Method components

In the DDNA approach the assignment matrix that maps OD flows to link flows is estimated using detailed GPS probe data. GPS probe data is used to estimate link travel times per time period (hourly for the project evaluation). A route set, i.e. all used routes in an OD pair, is also estimated based on the GPS probe data. The travel times and route sets are then combined to estimate route shares for all routes in all OD pairs using a logit-based route flow model. When assigning the OD demand to the road network, the route flows are propagated in the network according to the estimated travel times and finally the OD flows are adjusted so that the assigned link flows are as close as possible to the observed link flows. A detailed description of the OD estimation based on DDNA can be found in [9].

Since a significant part of the demand in the test area start and/or ends outside the area, country-wide demand from both SAMPERS and Telia must be aggregated to external zones of the test area. The demand that either start and/or end outside the test area is mapped to external OD pairs using a simple shortest path calculation. Further preprocessing that must be made for the analysis include, map matching of GPS probe data (provided by INRIX), map matching of link count sensors and mapping between the different zonings used for SAMPERS and Telia OD flows.

5.5 Evaluation

The DDNA-based OD estimation is evaluated for the Stockholm test network 2019-10-10 05:00-08:00 with 360 000 OD pairs, where 50 000 OD pairs are external. 80% of sensors (360 sensors) are selected randomly for training (estimating OD) and 20 % of sensors (90 sensors) are used for

test (validating OD). The training and validation data corresponds to 873 and 219 hourly link flow measurements, respectively.

Figure 63 shows the estimated link flows per hour based on DDNA. Figure 64 shows the spatial distribution of link flow errors. Figure 65 and Figure 66 shows examples of scatter plots with measured link flows plotted against estimated link flows for the training sensors and the validation sensors, respectively. Figure 67 shows a scatter plot of the the estimated link flows for two different selections of training sensors. Figure 68 and Figure 69 shows boxplots of the RMSE for the estimated OD and the SAMPERS OD when assigned to the network using DDNA for training sensors and validation sensors, respectively.

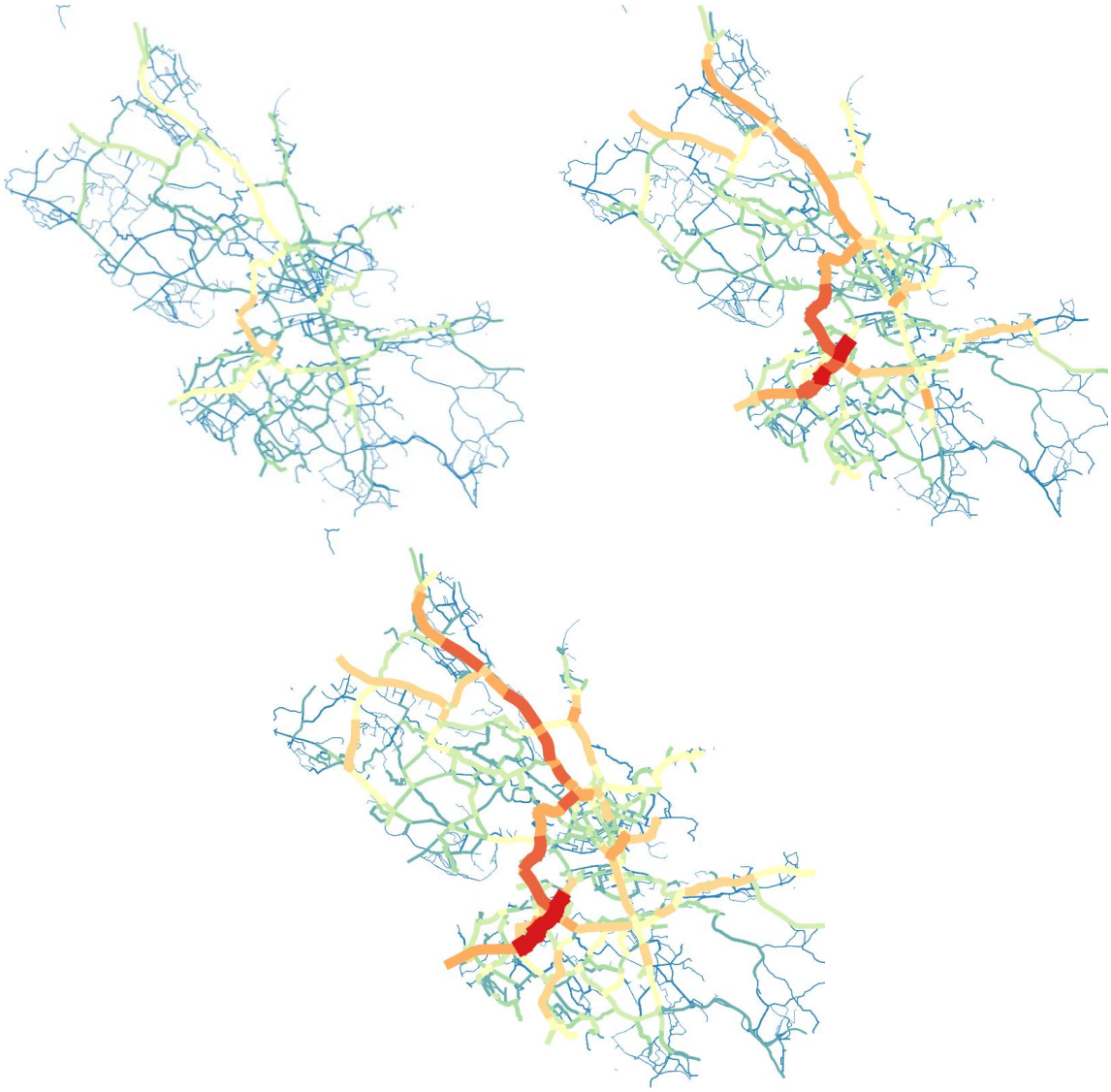


Figure 63 Link flow estimations based on DDNA 05.00-06.00 (top left), 06.00-07.00 (top right) and 07.00-08.00 (bottom).

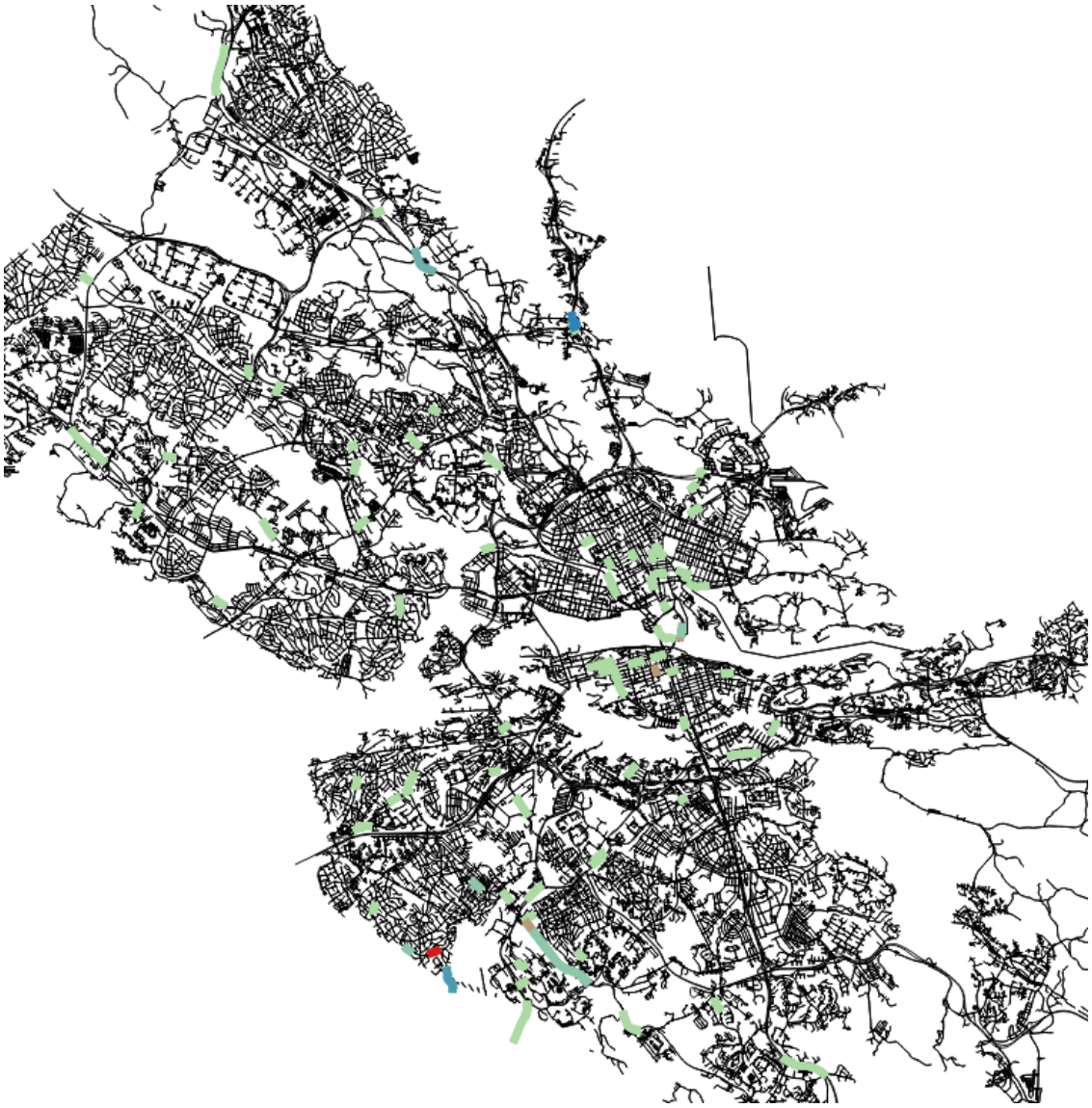


Figure 64 Spatial distribution of flow estimation errors.

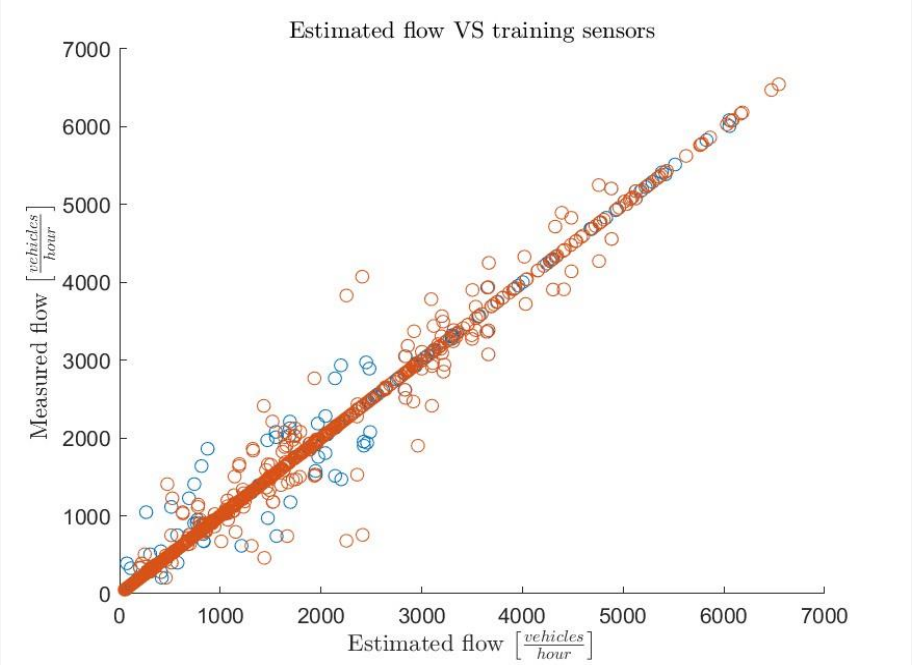


Figure 65 Scatter plot with estimated VS measured link flows for the training sensors.

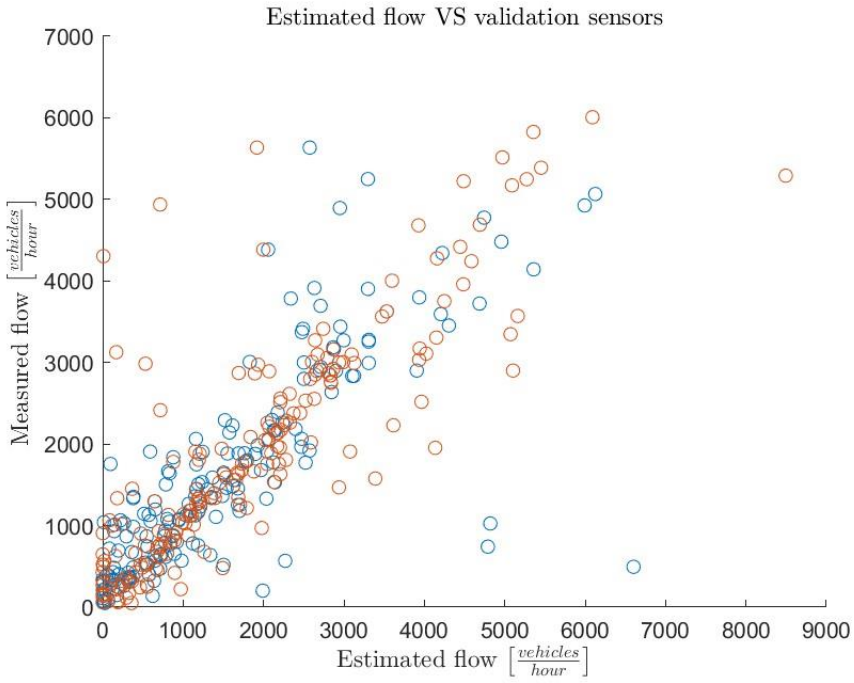


Figure 66 Scatter plot with estimated VS measured link flows for the validation sensors

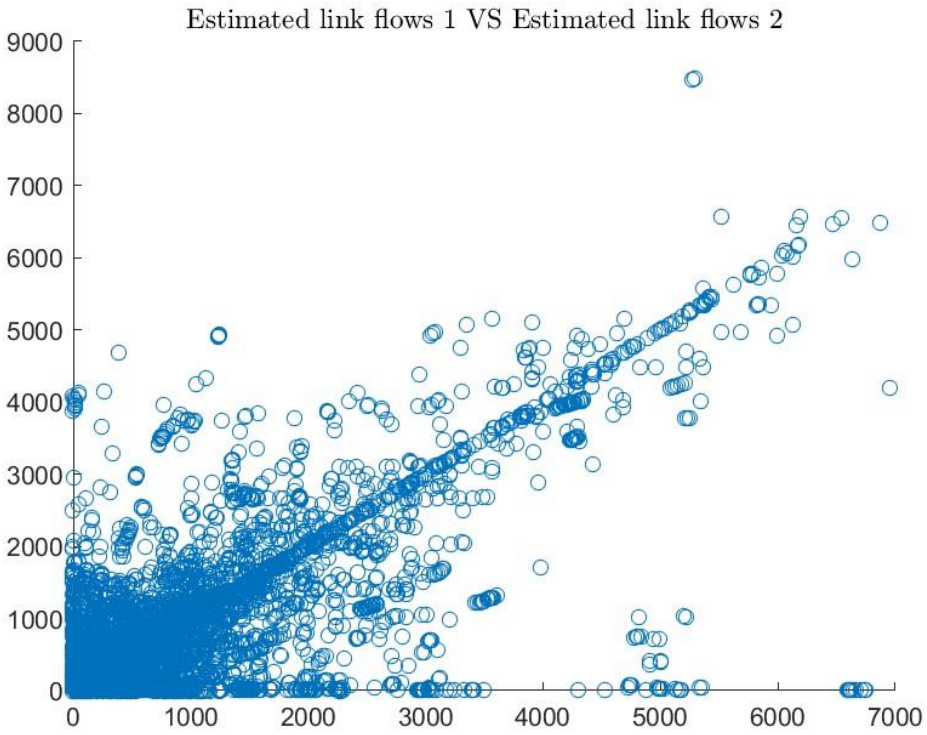


Figure 67 Scatter plot of estimated link flows for different choices of training sensors.

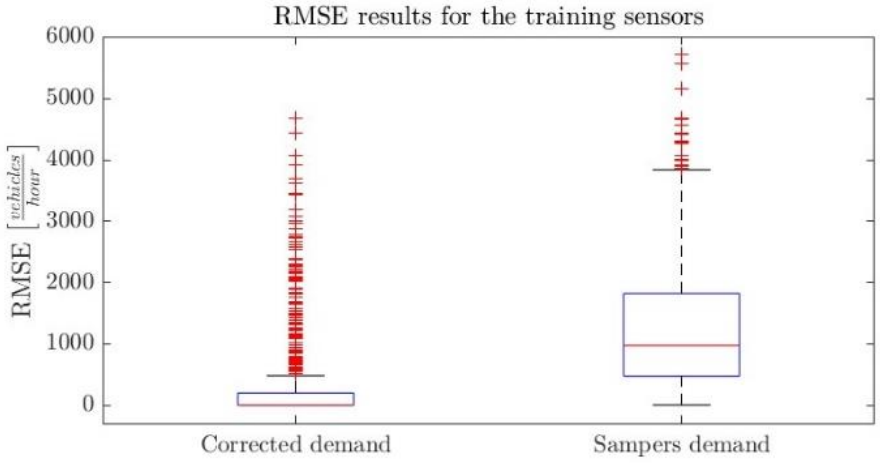


Figure 68 Box plot of RMSE for the estimated OD (left) and SAMPERS OD assigned using DDNA for the training sensors.

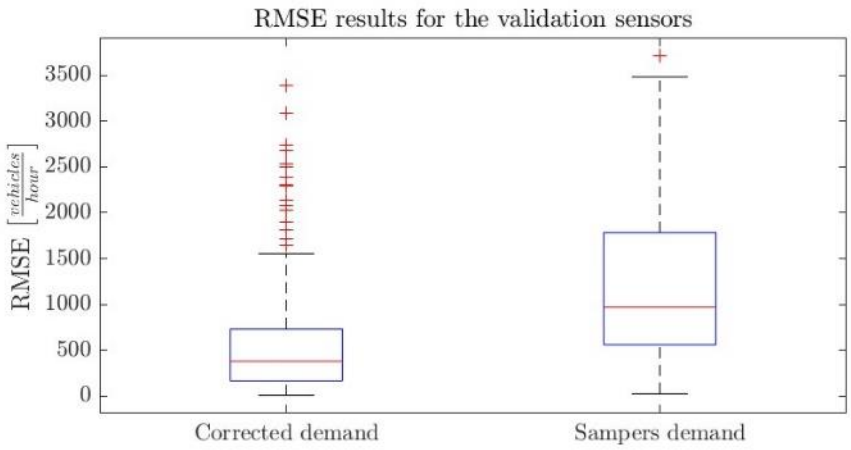


Figure 69 Box plot of RMSE for the estimated OD (left) and SAMPERS OD (right) assigned using DDNA for the validation sensors.

The results indicate that it is possible to estimate the OD such that the link flow estimates, when assigned using DDNA, matches the measured link flows better, both for the training sensors and the validation sensors. The results also indicate that the estimation models introduce large variance, i.e. the result is highly dependent on which sensors that are chosen for training the models. More work is needed to evaluate the method for longer periods of time and analyse how the different data sources should be combined for robust and accurate estimation of OD matrices.

6. DISCUSSION

6.1 Use of GPS data and survey

Among the survey respondents, the general knowledge and experience of using GPS data were low. Another sample of respondents might have provided different answers regarding their interests in further development. For future studies, it would be valuable to include a targeted group with more experience in using GPS data. Would the results be similar, with the same main barriers hindering the use of GPS data? Or could alternative sources be used to overcome the presented barriers, such as the lack of developed methods, missing processes and routines for using GPS data, and the cost associated with data?

Throughout the course of the study, the penetration rate poses a problem, especially on less urban roads. The trade-off between precision and having sufficient data is an important consideration when dealing with GPS data. The Nyköping study of a smaller town clearly showed that applying the data results in a sample that is too small for use, unless the data is aggregated into sufficiently large time periods. On the other hand, it is essential to acknowledge that the penetration rate has changed since 2019, opening up new possibilities for applications with newer datasets, especially on roads and in areas with higher penetration rates. When the entire day's data is aggregated, the ability to utilize the fine details available in the GPS dataset while detecting events is lost. For example, the disturbances caused by natural fluctuations during mornings, mid-days, and evenings are no longer discernible at the aggregated level.

As the Nyköping example shows, on the aggregated level, the GPS data can still be very useful even for the roads with relatively low penetration rates. For example, the shares of traffic during the peak hours, or across the days of the week, can be used to support modelling.

6.2 Event detection

In the sense of analysing and handling the changes in the road network day to day, it is important to acknowledge that both predictable and unpredictable events can be the cause of a disturbance of the system. Identifying a single factor and its effect is not always straightforward. However, by gaining understanding of the factors behind these events and their effects, the robustness of the transport system can increase, and possible negative impacts from the events can be reduced.

Road works, incidents, and large social events have an immediate impact on the road network. They hinder traffic flow, reduce capacity, and may even cause complete road closures for a limited time. These factors were taken into consideration when investigating the differences detected from the baseline, as mentioned in section 4.3. Reported incidents, as well as major road constructions affecting traffic, were cross-referenced with the detected events in Stockholm. For the Norrköping area, information was gathered from newspaper reports regarding incidents and disturbances since incident report data was not available for the Norrköping geographical area.

In this project, weather was not taken into account when investigating the case studies. However, similar to incidents and roadworks, various weather conditions can impact drivers and alter their driving behavior, subsequently affecting traffic flow. Weather can also be a factor that increases the likelihood of incidents. The influence of weather on traffic conditions, both during the calculation of the average situation and during the occurrence of the detected events, was not within the scope of this project.

As traffic demand naturally fluctuates for various reasons, there are days with higher and lower traffic volumes. The difference based on the day of the week can be significant, as observed in the study in Nyköping. Therefore, day-to-day fluctuations are expected, making direct comparisons

against a mean a complex task. A potential future method could involve clustering days together to identify natural differences, thus ensuring that an event or road segment that deviates from the norm is not merely a part of normal variation.

The case study events exhibit differences across several dimensions, including variations in time, location, and the nature of the events themselves. Significant events such as Lidingöloppet are expected to significantly influence mobility patterns in the car network. In contrast, the events like the climate demonstration and the bike carnival may attract participants who use alternative modes of transport or have shorter travel distances and might not affect road traffic flows as much. Furthermore, the central urban location amplifies the complexity of the study, making it challenging to pinpoint the reasons for changes in mobility patterns.

Furthermore, extending the dataset's time period would enhance the study's reliability, as a little over a month of data is a limited time frame for event detection.

Policy interventions, such as road construction or adding a new public transport service could also be analyzed using a similar framework. Time period extension would be especially important for such studies because this would allow to compare the situation "before" and "after" the policy intervention.

6.3 Link flow estimation

Both the local link flow estimation methods as well as the methods based on DDNA produces link flow estimates that explains a lot of variance in the data. If OD and route flows are also of interest, apart from link flows, it makes sense to utilise a method with consistent OD and link flow estimates. It should be noted that simple scaling of GPS probe data can also work for several applications where the prediction accuracy is not critical, however, machine learning approaches can improve the results if many link flow measurements are available.

The results indicate that it is easier to predict link flows in time (for a specific link) than in space, especially if the links are not part of the same major route. This is also why it is important to utilise training and validation datasets with different link flow measurements when different methods are evaluated. More detailed feature engineering will most likely improve the link flow predictions, especially for predictions in time.

6.4 OD estimation

The dynamic OD estimation problem for high-resolution zones similar to what we have used in this project is very challenging and it is difficult to find ground truth data. Mobile network data has enough temporal resolution to estimate continuous hourly OD matrices, but the GPS probe data used in this project has a too small penetration rate to work for OD estimation even on daily basis. With long timer periods it is possible to aggregate GPS probe data so that you can capture long-term mobility patterns.

The spatial resolution of the zoning is important for the characteristics of the OD estimation problem. With large zones it is possible to aggregate more GPS probe data and also reduce effects of limited spatial resolution of mobile network data, especially in rural areas. However, larger zoning also creates problems with internal zone flows that are not estimated in a good way. In this project we have worked with multiple zonings to utilize the different characteristics of the available data sources, but this needs further analysis for understanding what the effects will be on OD estimation accuracy.

7. CONCLUSIONS AND FUTURE WORK

In the project we have shown that detailed GPS probe data can be used for 1) direct link flow estimation, 2) identifying unusual events in traffic and 3) estimating an assignment matrix for mapping between OD and link flows.

The methods for predicting link flow measurements using GPS probe data can explain a large part of the variation in the data in hourly level, but the errors are relatively large, especially for links with few observations. Methods based on machine learning can improve the link flow prediction accuracy, but further work is needed to analyse which combination of features that should be used and which type of model that will give the best bias-variance trade-off when predicting link flows also for links without link flow measurements.

Direct scaling of OD flows is challenging for continuous hourly estimation, unless the spatial aggregation is very large. However, by aggregating data for longer time periods it is possible to get reasonable OD flows also from GPS probe data, at least when aggregated to zone flows.

GPS probe data can be used to detect events in the traffic system and utilising large-scale mobility patterns can improve the detection compared to only using local estimates of link counts. However, events with limited extension in time that occurs on roads with small flows without major changes in large-scale mobility patterns are hard to detect using only GPS probe data.

Estimating an assignment matrix using GPS data enables consistent and continuous OD and link flow estimation. The accuracy of the link flow estimates for DDNA-based link flow estimates are similar to the accuracy when using local scaling methods based on machine learning, with the advantage that we get an understanding of OD and route flows at the same time. Estimating the assignment matrix based on GPS probe data is challenging and more work is needed to evaluate both the assignment matrix, the link flows and the OD flows for larger time periods and new geographical areas.

When we analyse the spatial and temporal characteristics of GPS probe data and mobile network data, the data sources are complementing each other with higher spatial resolution of the GPS probe data and temporally higher resolution for the mobile network data. More work is needed to determine ways to combine the data sources efficiently for improving the results in both link and OD flow estimation.

Future work includes evaluation of the consistent OD and link flow estimation based on DDNA for a larger time period and new geographical areas as well as reducing the computation time for estimating the OD and link flows for large networks. To enable estimations for longer time periods (several years), it can also be of interest to estimate the assignment matrix without using detailed GPS probe data, which can be challenging to collect for longer time periods. Clustering of long time series can be interesting for automatic identification of traffic supply, demand and behavior changes while extending the OD matrix estimation to multiple vehicle classes is interesting from several planning perspectives.

Future work also includes more detailed analysis of how the methods can be applied for analysis of the impact of the events and incidents to analyze the effect of policy interventions, such as construction of new roads or bike highways, or adding a new public transport line. The comparison “before” and “after” the intervention can help to quantify the impact of these interventions, in terms of, for example, reduction or increase in car traffic and CO2 emissions.

REFERENCES

[1] S. Kim och B. Coifman, ”Comparing INRIX speed data against concurrent loop detector stations over several months”, *Transp. Res. Part C Emerg. Technol.*, vol. 49, s. 59–72, dec. 2014, doi: 10.1016/j.trc.2014.10.002.

[2] Y. O. Adu-Gyamfi, A. Sharma, S. Knickerbocker, N. Hawkins, och M. Jackson, ”Framework for Evaluating the Reliability of Wide-Area Probe Data”, *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2643, nr 1, s. 93–104, jan. 2017, doi: 10.3141/2643-11.

[3] V. Ahsani, M. Amin-Naseri, S. Knickerbocker, och A. Sharma, ”Quantitative analysis of probe data characteristics: Coverage, speed bias and congestion detection precision”, *J. Intell. Transp. Syst.*, vol. 23, nr 2, s. 103–119, mar. 2019, doi: 10.1080/15472450.2018.1502667.

[4] P. Sekuła, N. Marković, Z. V. Laan, och K. F. Sadabadi, ”Estimating Historical Hourly Traffic Volumes via Machine Learning and Vehicle Probe Data: A Maryland Case Study”, *ArXiv171100721 Cs Stat*, okt. 2018, Åtkomstdatum: dec. 22, 2020. [Online]. Tillgänglig vid: http://arxiv.org/abs/1711.00721.

[5] N. Tsanakas, J. Ekström, D. Gundlegård, J. Olstam, och C. Rydergren, ”Data-driven network loading”, Transportmetrica B: Transport Dynamics, s. 1–29, nov. 2020, doi: 10.1080/21680566.2020.1847213.

[6] Ahlberg, J., Danielsson, A., Drageryd, L., Gundlegård, D., Ramsey, J., Sjöholm, A., & Sjöstrand, S. (2021). Probedata: förstudie kring användning av gps-baserad probedata för skattning av hastigheter, länkflöden och ruttval (No. TRV 2019/98384). Retrieved from https://fudinfo.trafikverket.se/fudinfoexternwebb/Publikationer/Publikationer_004501_004600/

[7] Telia Sverige AB. (2023) Telia crowd insights. [Online]. Available: https://business.teliacompany.com/crowd-insights

[8] K. Ågren, P. Bjelkmar, and E. Allison, “The use of anonymized and aggregated telecom mobility data by a public health agency during the COVID-19 pandemic: Learnings from both the operator and agency perspective,” Data & Policy, vol. 3, p. e17, 2021

[9] N. Tsanakas, D. Gundlegård & C. Rydergren (2023) O–D matrix estimation based on data-driven network assignment, Transportmetrica B: Transport Dynamics, 11:1, 376-407, DOI: 10.1080/21680566.2022.2080128

APPENDIX

This appendix provides a more detailed description of the mathematical notation used in the machine learning models for local estimation of link flows. The following features are available in the dataset:

- "segment_row_id": identity of the link (categorical data)
- "road_class": type of road according to openstreetmap, i.e., 1 = highway, 6 = small residential road (categorical data)
- "sensor_id": identity of the sensor (categorical data)
- "sensor_type": type of sensor, medy = temporary link count, mcs = fixed radar station (categorical data)
- "dow": day of week, i.e., 0 = sunday, 6 = saturday, extracted from timestamp (categorical data)
- "hod": hour of day, extracted from timestamp (categorical data)
- "probe_flow": number of GPS probe vehicles that have passed the link during one hour (numerical data)
- "measured_flow": total number of vehicles per hour measured by the sensor (numerical data, target variable)

Note that "probe_flow" and "measured_flow" are numerical data and all others are categorical data.

The fraction of cars equipped with GPS sensor is called the penetration rate. One very simple estimator, which we will use as a baseline, is to estimate the traffic flow as estimated_flow = "probe_flow" / penetration_rate. In the machine learning models, only the features "road class", "dow", "hod" and "probe flow" are used.

In this project, both linear models and generalized linear models to are used to construct regression models that estimate the measured traffic flow ("measured_flow") based on multiple input features.

Linear model

Let n be the number of examples in the (augmented) training set:

$$\mathcal{S}_{\text{train}} = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$$

where $x^{(i)} \in \mathbb{R}^{d+1}$ is the ith input example consisting of an intercept term $x_0^{(i)} = 1$ and all other features (including the interactions terms) included in the model and $y^{(i)} \in \mathbb{R}$ represents the target (dependent) variables. We consider the following $d = 11$ input features: "dow", "hod", "road_class", "probe_flow", all possible interactions among them (i.e., "dow"×"hod", "dow"×"road_class", "hod"×"road_class", "dow"×"probe_flow", "hod"×"probe_flow", "road_class"×"probe_flow"), and "probe_flow^2".

Let $X \in \mathbb{R}^{n\times(d+1)}$ be the **design (model/regressor) matrix**, with the following structure:

$$X \triangleq \begin{bmatrix} - & (x^{(1)})^T & - \\ - & (x^{(2)})^T & - \\ & \vdots & \\ - & (x^{(n)})^T & - \end{bmatrix}.$$

We assume the following linear model

$$X\theta = \begin{bmatrix} - & (x^{(1)})^T & - \\ - & (x^{(2)})^T & - \\ & \vdots & \\ - & (x^{(n)})^T & - \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix} \approx \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix} \triangleq y$$

where $\theta \in \mathbb{R}^{(d+1)}$ are the model's parameters.

Linear regression (LS fit)

In the linear regression we define the loss function as the sum of the *squares* of the residuals, i.e.,

$$J(\theta) = \|X\theta - y\|^2,$$

and estimate the model's parameters (train the model) via ordinary least squares (LS), i.e.,

$$\theta^\star_{\text{LS}} = \arg\min_\theta J(\theta) = \arg\min_\theta \|X\theta - y\|^2$$

If X is **full rank**[1] then Gram matrix X^TX is invertible, and $\theta^\star_{\text{LS}}$ is **unique** and given by

$$\theta^\star_{\text{LS}} = (X^TX)^{-1}X^Ty$$

where $X^\dagger = (X^TX)^{-1}X^T$ is the Moore-Penrose **left (pseudo)inverse**[2] of X and **Gauss-Markov theorem** ensures that $\theta^\star_{\text{LS}}$ is a **BLUE estimator** for the model parameter θ.

[1] Skinny matrix X is full rank if its **columns (features) are linearly independent**; if so, $\text{rank}(X) = d + 1$.
[2] Note that $X^\dagger X = I$ but $XX^\dagger \neq I$

If X is not full rank (i.e., $\text{rank}(X) < d+1$) then Gram matrix $X^T X$ is singular, and there are multiple LS solutions $\theta^\star_{\text{LS}}$. The least-norm among them is also given by

$$\theta^\star_{\text{LN-LS}} = X^\dagger y$$

where $X^\dagger$ is still the left (pseudo)inverse of X (but it is computed differently via SVD) and $\theta^\star_{\text{LN-LS}}$ is still a **BLUE estimator** for the model parameter θ.

Poisson regression

We model the target variables (i.e., "measured_flow") as Poisson random variables

$$Y|x;\theta \sim \text{Poisson}(\lambda), \tag{1}$$

and assume that parameter λ is related to the input via[3]

$$\lambda = \exp(\theta^T x). \tag{2}$$

The maximum likelihood estimate of the model parameter θ is given by:

$$\theta^\star_{\text{Poisson}} = \arg\max_\theta \overbrace{p\left(y^{(1)},\ldots,y^{(n)}|x^{(1)},\ldots,x^{(n)};\theta\right)}^{\triangleq L(\theta)\quad \text{(conditional) likelihood}} = \arg\max_\theta \prod_{i=1}^n p\left(y^{(i)}|x^{(i)};\theta\right),$$

i.e., the value of θ that makes the observed data set most likely.

Instead of maximizing the (conditional) likelihood $L(\theta)$ it is computationally easier and mathematically equivalent to maximize its logarithm (i.e., the log-likelihood)

$$l(\theta) \triangleq \log L(\theta) = \sum_{i=1}^n \log p\left(y^{(i)}|x^{(i)};\theta\right),$$

thus

$$\theta^\star_{\text{Poisson}} = \arg\max_\theta l(\theta). \tag{3}$$

The optimization problem (3) is convex and, therefore, it be solved easily via the gradient ascent method, where the gradient of the log-likelihood is given by

$$\nabla_\theta l(\theta) = \sum_{i=1}^n (y^{(i)} - e^{\theta^T x^{(i)}}) x^{(i)} \tag{4}$$

Once $\theta^\star_{\text{Poisson}}$ is found, the Poisson regression model predicts the target variables (i.e., "measured_flow") as

$$\hat{y} = \mathbb{E}[Y \mid x;\theta^\star_{\text{Poisson}}] \overset{(1)}{=} \lambda \overset{(2)}{=} \exp\left({\theta^\star_{\text{Poisson}}}^T x\right). \tag{5}$$

[3]It can be shown that Poisson distribution is a member of the exponential family with natural parameter $\eta = \log\lambda$, hence $\lambda = e^\eta$.

Least squares alternatives: customized penalty functions

In many practical applications, we are not interested in minimizing the sum of the *squares* of the residuals, then LS fit does not capture the desired optimization criteria. An alternative but closely related view (or motivation) is that LS is highly sensitive to outliers and, therefore, it is preferable to use robust regression methods instead of simple LS.

Robust regression via l_1-loss function

This is similar to LS but instead of minimizing the *sum of the squares* of the residuals it minimizes the *sum of the absolute values of the residuals*, i.e., the l_1-norm of the residuals

$$\theta^{\star}_{l_1} = \arg\min_{\theta} \|X\theta - y\|_1 \tag{6}$$

and therefore it is also known as least absolute deviation (LAD), least absolute errors (LAE), least absolute residuals (LAR), or least absolute values (LAV).

An alternative probabilistic interpretation (or motivation) is the maximum likelihood model estimation assuming that the modelling errors follow a Laplace (instead of Gaussian) distribution. Another closely related idea is the quantile regression which estimates the *conditional median* (or other quantiles) of the response variable (instead of conditional mean used in LS).

The loss function in (6) is not smooth, therefore it cannot be solved via straightforward gradient descent methods. However, it can be cast as a linear program and solved using CVX (see this Youtube) or "linprog" function in Matlab (see MATLAB Answers, Constrained minimum L1-norm solutions of linear equations, Github code; alternative methods can be also found: here and here).

Robust regression via weighted-l_1-loss function

With a simple scaling we can also turn the regression via l_1-loss function into mean absolute percentage error (MAPE) minimization, also known as mean absolute percentage deviation (MAPD) minimization.

$$\theta^{\star}_{\text{weighted}-l_1} = \arg\min_{\theta} \| \operatorname{diag}(y)^{-1}(X\theta - y)\|_1 = \arg\min_{\theta} \| \operatorname{diag}(y)^{-1}X\theta - 1_n\|_1 \tag{7}$$

which can be solved using the same liner programming based methodology as in problem (6).

Note that the problem become ill-posed when some entries in y are zero. Furthermore, it over-emphasize the errors in the training examples with small values of measured flow $y^{(i)}$ (e.g., small residential roads). A solution to this problem is to perform a partial scaling which excludes the examples with $y^{(i)}$ smaller than a given threshold (where the threshold usually is optimized according to a specified performance criterion).