# Using sparse phone location data to estimate traffic volume and traffic speed

Patrick Andersson
David Fagrell
Tim Zetterquist

Advisor: Jorge Gil

Associate Professor and Head of Division Urban Design and Planning,

Architecture and Civil Engineering at Chalmers

Docent at Department of Computer Science and Engineering at Gothenburg University

November 20, 2023

**Abstract**

This project explores the possibility of estimating traffic flow and traffic speed on 10 road segments in and around Gothenburg, Sweden using a sparse commercial dataset with mobile phone location data spanning seven months in 2019. The hypotheses are $H_1$: Higher quality trajectory data, meaning a higher rate of data points, gives estimates closer to ground truth. $H_2$; This type of dataset can be used to estimate the annual average daily traffic on different types of roads. $H_3$: This type of dataset can be used to estimate hourly traffic states, such as free- flow and congestion. Data on hourly mean speed and annual average daily traffic from the Swedish government agency Trafikverket was collected as ground truth. After exploring these topics, the conclusion was made that there is no significant difference when estimating speed between high- and low-rate data points. The linear regression between dataset car count and the annual average daily traffic is significant but with a large RMSE of more than 7000 cars, and both Spearman and Pearson correlation coefficients are strong and significant. Estimating hourly speed and flow states had an RMSE of 10.4km/h for roads with congestion traffic patterns and an RMSE of 14.7 km/h for roads with free-flow traffic patterns.

# Contents

# 1 Introduction

Efficient urban mobility is a fundamental goal for any city planner of today. With roads in many urban centers congesting on a daily basis, delaying people and goods, and spewing emissions into the air, intelligent traffic management has quickly become a top priority. Traffic- and city planners rely on many sources of data to perform their jobs. In Sweden, it is the government agency Trafikverket (2022) that measures, collects, and distributes data about traffic on state roads, and smaller roads are handled by municipality authorities. The long-term statistics are derived from local and time-limited roadside measurements, and short to real-time analysis is already today done by aggregating data from multiple sources, including Global Positioning System (GPS) data from mobile phones and images from traffic monitoring cameras.

The infrastructure required for the statistics is quite expensive and is only done locally both spatially and temporally. The question is then, can we use only sparse GPS data to estimate traffic flows? In that case, a much more temporally present and spatially holistic picture of traffic conditions can be drawn, giving traffic managers another tool to make the best decisions they can.

The future of intelligent traffic management is bright as more and more vehicles are connected cars (Coppola & Morisio, 2016) in a paradigm of Internet of Vehicles (Sharma & Kaushik, 2019) that will produce and deliver high-quality location- and sensor data in a timely manner. These technologies are, however, not widely implemented yet and the data that does exist is proprietary.

This master's program course project will investigate the possibility of using a small commercial dataset to estimate traffic flow and speed on selected road sections in the city of Gothenburg Sweden. To guide our exploration we have come up with three hypotheses to test:

**Hypothesis 1** *Higher quality trajectory data, meaning a higher rate of data points, gives estimates closer to ground truth.*

**Hypothesis 2** *This type of dataset*[1] *can be used to estimate the annual average daily traffic*[2] *on different types of roads.*

**Hypothesis 3** *This type of dataset*[3] *can be used to estimate hourly traffic states, such as free-flow and congestion.*

The data rate's purpose in Hypothesis 1 is to provide a method for evaluation of the data quality. The time and the navigational coordinates recorded are used to compute informative measures such as velocity, acceleration, and other measurements useful for analyzing traffic flow. Therein lies the assumption; if the data points have a large duration between the recorded times, estimating the trajectory's navigational route, velocity, acceleration, and further deducted measurements are faulty. As the duration between the recorded data points increases, so do the errors in the estimation of deduced measurements. By applying this rate, one can hypothetically conclude the quality of the data.

Section 2 will define what we mean by traffic flow and describe how it is currently measured. Section 3 will cover current research on estimating traffic flow and speed using GPS data. Section 4 will present the data in this study and describe it, and section 4.1 will describe the process of cleaning and pre-processing the data, section 4.2 will describe the data used as ground truth, before section 5 describes our methods. Results are presented in section 6, discussed in section 7, and followed by our conclusions in section 8.

[1] see section 4 for a description of the dataset.

[2] Specifically the statistic from Trafikverket called ÅDT, which will be presented and explained in section 2.1.

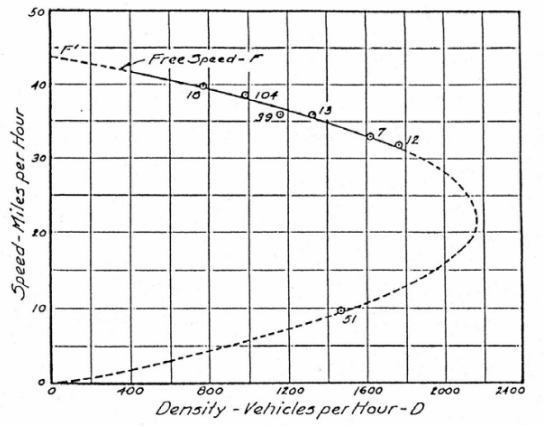
[3] see footnote 1

# 2 Traffic flow and speed

Traffic flow is an extremely complex spatiotemporal process involving many non-linear interacting agents. Greenshields et al. (1934) published the first empirical research of traffic flow on highways in the Proceedings of the 13th Annual Meeting of the Highway Research Board, where the researchers for the first time used photography to count vehicles and estimate their speed, and also applied mathematics to traffic flow. The research resulted in a simple linear equation of the observed speed-density relationship and also what has come to be known as the *fundamental diagram*, Greenshield's original diagram is shown in figure 1. The term *traffic flow* did not exist in traffic research at this time, so Greenshield used two different measures he called density - one that counted per hour and another that counted vehicles per mile of pavement (Kühne, 2011).

**Figure 1**

*Greenshields' original speed-density diagram from Greenshields (1934).*

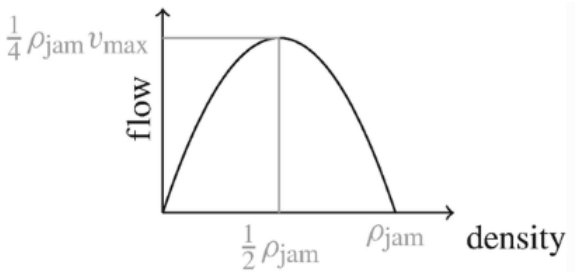



Since the time of Greenshields' research many shapes of the fundamental diagram have been developed, exploring the relationships between the variables flow, density, and speed (Kessels, 2019), where traffic flow is defined as the rate of the number of vehicles per time unit (*e.g.* $vehicles/hour$) and density as the rate of the number of vehicles per space unit (*e.g.* $vehicles/km$) (Gkountouna et al., 2020). The flow-density variant might be most illustrious for the purposes of this review, see figure 2.

The flow-density fundamental diagram illustrates a phenomenon of traffic jams, that the flow *decreases* when density has reached a certain threshold, rather than reaching a maximum flow that is maintained. This maximum flow is often referred to as the road capacity.

**Figure 2**

*The flow-density diagram, from Kessels (2019), where $\rho$ is density, $v$ is speed.*



A number of different techniques are used in order to get a sense of traffic flows. According to Tu et al. (2021), a currently common theory for describing it is the Kerner three-phase theory (Kerner et al., 2007) which divides traffic flow into three states. The *free-flow state* where there are fewer vehicles on the road and lower traffic density allows vehicles to easily reach the designed speed. The *steady flow state* is where traffic flow reaches the road capacity but the head-space between vehicles is shortened leading to decreased ability to operate and change lanes. Under a *congested state*, the vehicles on the road follow the vehicles ahead, make frequent stops, and move forward slowly which creates unstable driving conditions.

## 2.1 The annual average daily traffic statistics

Trafikverket is responsible for statistics concerning all state-owned roads in Sweden (*e.g.* all highways) and the annual average daily traffic statistics are used to make decisions that concern traffic planning, design of roads, consideration of environmental factors, and safety and maintenance of the road network. The method used for estimating this statistic is detailed in a report from Trafikverket (2015), where we learn that the traffic on segments of the larger road categories is measured every four years and that traffic on smaller roads is measured at least every 12 years. For each traffic segment, a sample selection survey is performed to choose a number of random measuring days that would give the yearly estimates a 95% confidence interval.

The measurements are mainly performed using two pneumatic tubes that cross a road a certain distance apart from each other[4] and that each sends an impulse when wheel axles roll over them. The system used by Trafikverket has been shown to classify 99.1% of all vehicles that pass the measuring point and is assumed to correctly identify vehicle category since the parameters *number of axles on road* ("antal axlar på väg") is correct 99.8% of the time and *wheelbase* ("axelavstånd") has a high precision (average deviation is 3.8cm). A few (approximately 50) segments are not suitable for this measuring equipment and are as such measured by other means (*i.e.*, Motorway Control Systems and permanent measuring points in the form of inductive loops).

These annual average daily traffic-statistics are regarded as robust estimates based on real and accurate measurements, and as such seen as ground truth by many decision-makers. One issue is however that the measurements are sparse both temporally and spatially, there is no way to get broad real-time insights. Another issue is that the measurements are quite expensive to perform, involving equipment, logistics, and field personnel. One hope in the field is that readily available GPS data from mobile phones can deliver the same insights and more.

# 3 Earlier research and the state of the art

All research regarding estimating traffic flow and speed from phone location data, that the authors of this report has read, first clean the data from obviously erroneous trajectory-segments and then apply some kind of map-matching technique to estimate the actual paths represented by the trajectory.

[4] 330cm is the distance used by Trafikverket, which is also the wheelbase that differentiates trucks from smaller vehicles.

## 3.1 Cleaning and repairing errors

The traffic prediction survey by Yuan & Li (2021) and the overview of trajectory data mining by Zheng (2015) both specify the need to handle missing data and outliers. Methods for outlier detection include creating anomaly scores from spatial distance measures or graph-based distance measures, and DBSCAN (Ester et al., 1996) clustering where points without clusters are regarded as outliers. Small errors are acceptable and can be handled by a *map-matching* procedure which will be discussed below, larger errors should be corrected using noise filtering methods such as a *mean filter* (Zheng, 2015).

Wang et al. (2013) lists and describe some higher level data cleaning filters that are generally applicable to trajectory data from urban environments. Their process involves removing:

- Unrealistic coordinates
- Segments with unrealistically high speed
- Segments with long distance
- Segments with long time

The thresholds for each filter is determined depending on the data and application, and also perhaps local domain knowledge on what constitutes a long distance or long time. At this point the segments are still straight line flight-paths between data points, perhaps cutting straight through buildings or across entire neighborhoods. To map segments to the road network a map-matching algorithm should be employed.

### 3.1.1 Map-matching

To accurately map trajectories to a road network is a complex task. A recent survey on map-matching methods (Chao et al., 2020) describes the history of the field as evolving from simple *geometric* methods to *topological* and *probabilistic* methods, to *advanced* methods involving advanced models such as the Kalman Filter (Zhao et al., 2003) to produce joint probability distributions. The trajectory data review by Zheng (2015) describe map-matching methods based on these categories, while Chao et al. (2020) goes on to describe four new classes of map-matching models based on the "core matching model": *similarity model, state-transition model, candidate-evolving model* and *scoring model*.

As can be inferred, map-matching algorithms vary greatly in complexity. Very recent research from Jiang et al. (2023) show very impressive results in matching low quality (very low sampling frequencies and noisy points) using multiple deep learning models that learn high-quality representations. They then mine for patterns in the latent space using a Gaussian mixture model (Murphy Kevin P., 2012) and incorporate these patterns in the loss function. A decoder followed by a joint optimization procedure then produces the final path. Compared to a Hidden Markov Model (Murphy Kevin P., 2012) and another deep learning method when dealing with really low-quality trajectories (with a sampling frequency of 90 seconds or more), their method is faster, more accurate, and more robust to a number of factors. In the end, the choice of map-matching method will depend on factors such as the quality of trajectory data, map density (many or few roads as candidate paths), acceptable error rates and resources available.

## 3.2 Estimating flow and speed

From section 2 we know that traffic flow classically is a function of speed and density, but that is obviously a simplified model that is constrained by a number of factors of the road segment that the function describes. With a data sample that is representative of the traffic one can do an

actual count of the trajectories and linearly extrapolate the count to the actual population, and as described below one could consider if some trajectories are ride-sharing or on the same bus.

Xing et al. (2019) use a clustering method to identify trajectories that are in the same vehicle to get an accurate estimate of the number of vehicles and also to extract information about which transportation mode a trajectory is using. Their dataset was cell tower data from all users of a single network operator in Nanjing, China, a dataset that is assumed to be representative of the entire city population. They then use a simple linear regression model to estimate actual flows. Their results indicate a strong correlation between their estimated flows and the actual measured flows. In a commercial dataset, this assumption is not as obvious and as such would be interesting to explore.

On the other hand if the dataset is not representative, flow must be derived from more factors than just a count of trajectories. Gkountouna et al. (2020) use a *segment archetype approach* where they cluster measured road segments into archetypes based on the latent features of traffic patterns derived from the actual speed- and flow time series and develop a flow estimation model for each archetype. They then classify (using the original cluster archetypes as training labels) each new non-measured segment as one of the archetypes based on the available speed time series for that segment and also spatial characteristics, and apply the flow estimation model corresponding to that segment archetype. This method bypasses the selection bias inherent in commercial phone-generated GPS data, at least somewhat, by embedding the relationship between actual speed and actual flow in the segment archetype model.

As with many other methods involved in the workflow of estimating traffic flow from GPS trajectory data, the choice of computation depends on the data available, the purpose of the investigation, resources available and previous steps of the workflow.

# 4 Data description and data handling

The dataset analyzed in this report was provided to the project group by the project supervisor. The original commercial dataset was bought by Chalmers University from two different data broker firms. First from a firm named Predicio[5] and later from Pickwell[6]. The data is gathered from a large number of apps from a large number of app categories where location data has been recorded either passively (*e.g.* in the background to record travel or exercise) or actively (*e.g.* "checking in" in locations). The location data is coupled with a device ID so to be GDPR-compliant the project groups received access to aggregated datasets without any device ID.

Within the original dataset, there were 5 billion points of raw location data with differing precision depending on the app used and the user preferences. Aggregated from this, our supervisor Jorge Gil provided this project group with a trajectory dataset consisting of two tables: one with approximately 2.5 million trajectories that intersect Gothenburg City and some of its surrounding areas, and one table consisting of all the approximately 13.5 million time-stamped data points inside that same area that belong to the trajectories.

The trajectories are stored as a sequence of coordinate x-y-pairs with a unique trajectory ID. Each trajectory also has aggregates for the number of data points it's made up of, the total Euclidean distance between the sequential points, and the average speed over that distance. Each data point has its own ID and also the corresponding trajectory ID that it is a part of.

[5] A firm that does not exist anymore.

[6] www.pickwell.co

## 4.1 Cleaning and pre-processing

For the purposes of this analysis, trajectories with only two recorded data points and an average velocity above 200 km/h or below 10 km/h were removed. These trajectories were visually inspected and assumed to be erroneous data considering that speed limits on Swedish roads never exceed 120 km/h and traveling with an average velocity above 200 km/h nears physical impossibility, as for the trajectories with a mean velocity of less than 10 km/h, they were assumed to be pedestrians. The reason for not deleting trajectories with more than two data points is that the high velocity can be the result of only a few erroneous segments in an otherwise fine trajectory.

Since this project is only interested in a few short segments of the road and not entire trajectory paths, a rudimentary map-matching based only on the heading of the trajectory intersecting the road section was employed after removing all trajectories that do not cross any of the investigated road sections. This process is described in section 5.

## 4.2 Validation data

To validate the estimations, data was collected from Trafikverket's (2023) website that hosts an interactive map. From this map, there is information on the annual average daily traffic, mean speed on an hourly basis, season variation, et cetera on a selection of roads. Multiple road sections were assessed to find ones that met the traffic pattern criteria, which was to either have a steady average speed over the day or have a significant drop in speed during the afternoon. Upon finding 10 suitable roads, the hourly mean speed, number of passing vehicles, and annual average daily traffic were extracted to use as ground truth.

For testing hypothesis 1 and 3 the selected day for each road closely aligned with the dataset year, 2019, and fell between June and December to coincide with the dataset's months. For testing hypothesis 2 the annual average daily traffic for 2019 was chosen if data was available, otherwise the closest year with data that came before was chosen, since data from 2020 and 2021 very well might be non-representative due to factors caused by the COVID-19 pandemic. The selected road sections together with their observed traffic pattern, their annual average daily traffic, the dataset car count, and the ground truth of the number of cars, one way, on the chosen day for hypothesis 1 and 3) are listed in table 1. The location of the road sections is shown in Figure 7 of Appendix A.

**Table 1**

*Selected road sections with their traffic pattern, annual average daily traffic, dataset car count, and ground truth for hypotheses 1 & 3.*

| Section id | Section name | Traffic pattern | AADT (n) | car count (n) | $H_1$ & $H_3$ truth (n) |
|---|---|---|---|---|---|
| 1 | E6 Myrstena | Congestion | 52,067 | 2584 | 27,534 |
| 2 | Agnesberg | Congestion | 32,637 | 2416 | 16,815 |
| 3 | Jonsered | Free-flow | 40,332 | 2063 | 21,955 |
| 4 | Torpamotet | Congestion | 63,285 | 3542 | 7,822 |
| 5 | Oscarsleden | Free-flow | 44,056 | 2171 | 24,640 |
| 6 | E6 Kallebäck | Congestion | 63,863 | 5292 | 33,852 |
| 7 | Riksväg 40 | Congestion | 60,315 | 3909 | 34,759 |
| 8 | Järnbrott | Free-flow | 62,193 | 2890 | 37,817 |
| 9 | Åbymotet | Free-flow | 71,505 | 5594 | 16,403 |
| 10 | Askim | Free-flow | 27,857 | 1481 | 18,640 |

# 5 Methods

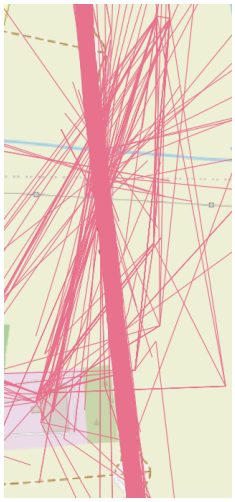
To count vehicles passing through a road section, first, a line geometry object for each road section was created using the free open-source software QGIS Geographic Information System (QGIS.org, 2023). Then a buffer polygon of 20 meters was created for each line using the Python (Van Rossum & Drake, 2009) library Shapely (Gillies et al., 2023) and the trajectories intersecting these polygons were extracted by means of a spatial join using the Python library GeoPandas (den Bossche et al., 2023). A very simple map-matching based only on direction was used to exclude trajectories that cross rather than travel along the road section. Allowed headings were the road heading $\pm 10$ degrees. See figure 3 for a comparison of before and after filtering on heading. The entire dataset was reduced to 23,400 trajectories with this filter, with 15,378 trajectories intersecting road sections with congested traffic patterns and 12,780 trajectories intersecting free-flow sections. The sum is greater since some original trajectories intersect more than one road section.

Once relevant trajectories for a road section are established the procedure differs depending on which hypothesis is to be tested.

**Figure 3**

*Before and after map-matching trajectories that cross road section 1 (E6 Myrstena).*

(a) *Unfiltered trajectories.*



(b) *Filtered trajectories.*

## 5.1 Hypothesis 1

This hypothesis states that trajectories with a higher rate of recorded data points provide data with higher quality than trajectories with a lower rate of recorded data points. For this, a ratio for every trajectory, *i*, is calculated as a method for determining the quality of a recorded trajectory. The ratio, *r*, is calculated as the number of recorded data points in a trajectory, *n*, divided by the temporal duration of the trajectory, *t*.

$$r_i = \frac{n_i}{t_i}$$

For instance, if a recorded trajectory has a duration of 600 seconds and the total number of recorded data points in the trajectory is 60. Then the ratio is 0.1 data points per second. I.e. there is one data point for every 10 seconds in the trajectory.

To test the difference in the ratios between erroneous data and accurate data, a two-sample $t$-test with a chosen significance level $\alpha = 0.05$ was performed (Ozdemir et al., 2018). Assuming that trajectory data with a mean velocity $\geq 120$ km/h contains erroneous data, we group the data into trajectories with a mean velocity $\geq 120$ km/h and mean velocity $< 120$ km/h. Also, for the purposes of the test, the trajectories with a mean velocity below 10 km/h were removed from the analysis. This is due to the data containing information from pedestrians and is irrelevant to the analysis of motor traffic on highways. The hypothesis tested is as follows:

> $H_0$: The mean data rate is equal between trajectories with mean velocity $< 120$ km/h and trajectories with mean velocity $\geq 120$ km/h.
>
> $H_1$: The mean of the number of data points/duration is not equal between trajectories with mean velocity $< 120$ km/h and trajectories with mean velocity $\geq 120$ km/h.

The trajectories were filtered as described in the first paragraph of section 5 to only keep those that intersect with a road section and seem to travel along it. These filtered trajectories were visualized in QGIS to ensure they corresponded to the actual roads.

Next, the data was sorted by the rate of data points over duration and split in half to see whether a high- or low rate was closer to the true mean speed for the chosen roads. The median value for all rates was 0.19, i.e., approximately one data point every five seconds. So the lower half had less or an equal amount of data points over the duration, while the upper half had more or equal amounts of data points over the duration.

The hourly mean velocity was plotted using Matplotlib (Hunter, 2007) for both rates with data from Trafikverket and the root mean squared error was calculated for all road sections, both individually and aggregated.

## 5.2 Hypothesis 2

Hypothesis 2 states that we should be able to estimate the annual average daily traffic on different types of roads. The basic approach is to count trajectories on certain road segments and compare the count to the official annual average daily traffic statistics from Trafikverket. From this, if the dataset is representative of city-wide traffic, we hope to extract a linear relationship that might be generalizable to other road segments.

Since the dataset isn't expected to be representative, one would optimally perform at least a partial replication of for example Xing et al. (2019), which is described in section 3. Due to a lack of actual measured data for training, and resource and time limitations in this project, this was not accomplished.

To test the correlation between the dataset vehicle count and actual annual average daily traffic, both a Spearman's rank correlation coefficient and a Pearson correlation coefficient were computed (Skiena, 2017).

## 5.3 Hypothesis 3

Hypothesis 3 seeks to determine if sparse and non-representative phone location data can accurately predict hourly traffic conditions, such as free-flow and congestion.

To filter trajectories from the GPS dataset that passed the specified road sections, the same procedure as in hypothesis 1 was used.

Additionally, all trajectories, whether indicating free-flow or congestion, were categorized by hour over the entire duration, and average speeds were calculated. This method was replicated for Trafikverket's data, enabling a side-by-side comparison, as illustrated in Figure 6.

# 6 Results

## 6.1 Hypothesis 1

The two-sample $t$-test showed that the differences in means of data rate between the two groups are statistically significant. Trajectories with a mean velocity $\geq$ 120 km/h had significantly lesser mean data rates than trajectories with a mean velocity $<$ 120 km/h; $t(518113) = 6.060$, $p < 0.001$.

**Table 2**

*$t$-test results for trajectories with a mean velocity of <120 km/h and >= 120 km/h*

| Group | n | Mean | Variance | Standard Deviation |
|---|---|---|---|---|
| <120 km/h | 466284 | 0,119 | 4848,174 | 69,629 |
| >= 120 km/h | 51831 | 0,116 | 339,129 | 18,415 |

By plotting the aggregated mean velocity data for both high and low ratios of trajectories below 120 km/h alongside the mean speed data from Trafikverket, we observed some differences in the estimations. For congested roads, the low ratio yielded an root mean squared error (RMSE) (Ozdemir et al., 2018) of 11.9 km/h, while the high ratio resulted in an RMSE of 8.5 km/h. Conversely, for free-flowing roads, the low ratio had an RMSE of 17.0 km/h, and the high ratio had an RMSE of 11.5 km/h. So, GPS data with a data point rate of more than one data point every five seconds is slightly more accurate than GPS data with a lower rate.

As depicted in figure 4, the GPS data catches the speed drop during the morning rush hour, but two hours earlier than the data from Trafikverket. However, the main difference is the fact that it doesn't go back up as much as the true data, but rather goes to about 75 km/h and then slowly decreases until the afternoon rush hour. Furthermore, the estimated speeds lie around 10 km/h below the actual speed throughout the day, and the low rate speed is lower than the high rate during almost all hours of the day for both free-flow and congested roads.

## 6.2 Hypothesis 2

To test whether the dataset car count significantly predicts the ground truth annual average daily traffic, a simple linear regression was performed (Ozdemir et al., 2018), see figure 5. The fitted regression model was:

$$\hat{\text{annual average}} \text{ daily traffic} = 22675.1914 + 9.1212x$$

The overall regression was statistically significant ($R^2 = 0.72, p = 0.002$). The RMSE was 7441.83, the smallest error was 1160.20, and the largest error was 13157.58.

To further explore the relationship between the dataset car count and annual average daily traffic, a Spearman's rank correlation coefficient as well as a Pearson correlation coefficient were computed. Both tests indicate a strong positive correlation.

**Figure 4**

*Mean speed for low ratio, high ratio, and true data on congested roads.*

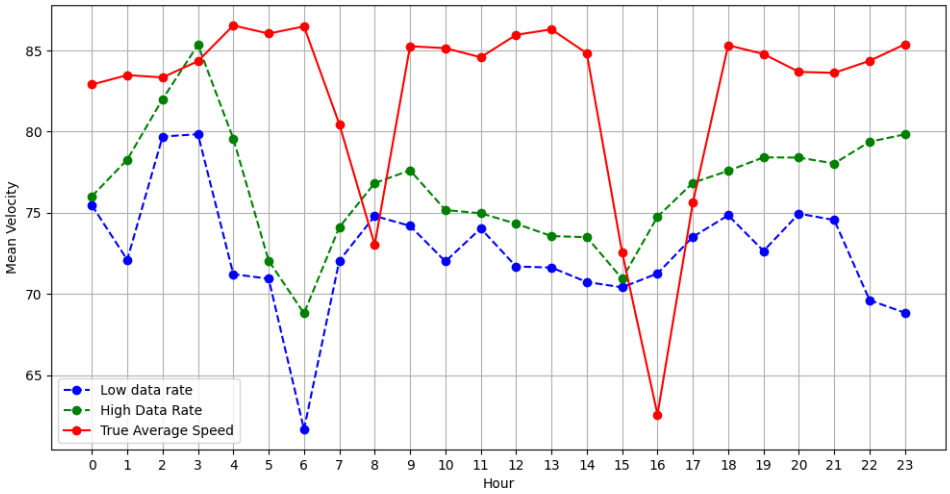
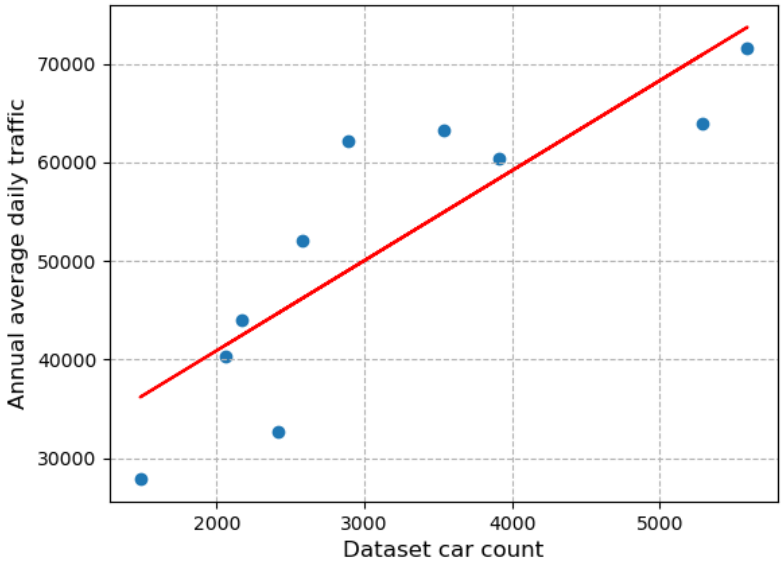



**Figure 5**

*Linear regression of dataset car count on annual average daily traffic*




Spearman's rank coefficient indicates a positive correlation between the two variables, $r(8) = .93, p < .001$, and the Pearson correlation coefficient also indicates a positive correlation $r(8) = .85, p = 0.002$.

The fact that Spearman's $r$ is larger than Person's $r$ indicates that the relationship is monotonic and non-linear.

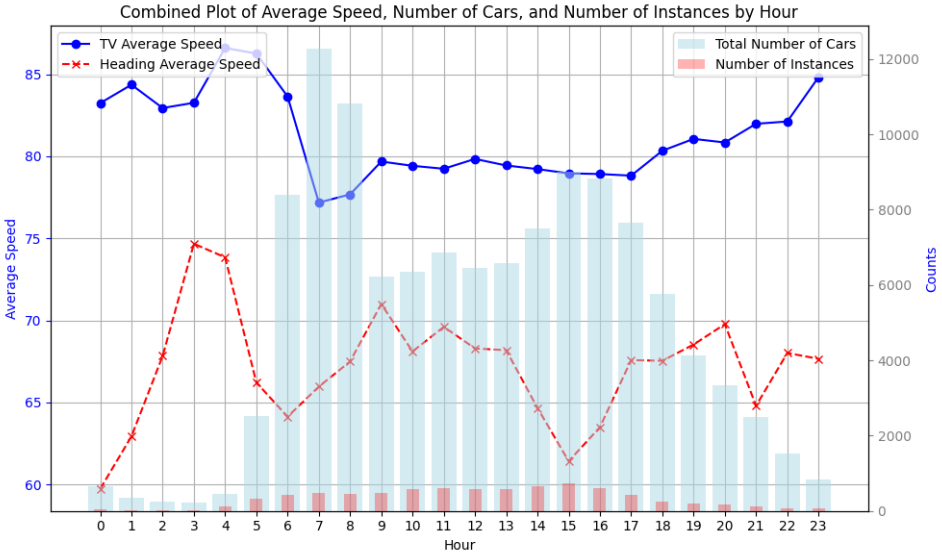
## 6.3 Hypothesis 3

When visually comparing the estimated mean speed and number of vehicles passing all road sections for the combined congested- and free-flow roads, they follow the general trends of the true data as seen in Figure 6, but lie about 10-15 km/h below the validation data. The GPS data had an RMSE of 10.4 km/h for the congested road and an RMSE of 14.7 km/h for the free-flow roads. A pattern appearing in both plots is that the estimated average speed follows a similar variation in speed, but occurs about one hour before the true major mean speed changes. The main difference is greater fluctuations throughout the day in both plots, especially during nighttime, and the estimated mean speed in the free-flow roads has a major drop during afternoon rush hour that is not present for the true data while the GPS data doesn't catch the afternoon rush hour for the congested roads.

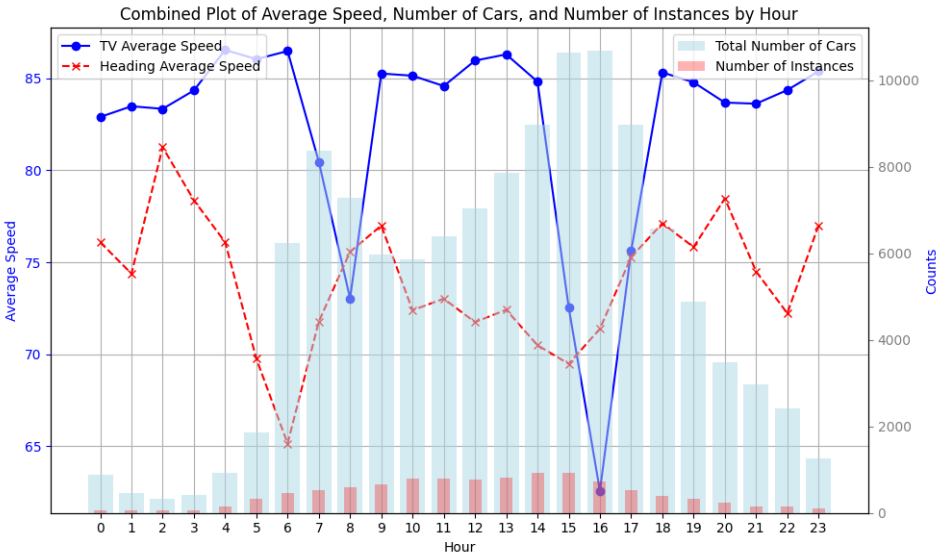
**Figure 6**

*Hourly mean speed of true and estimated values*

(a) *Afternoon free-flow*




(b) *Afternoon congested*



# 7 Discussion

## 7.1 Hypothesis 1

After conducting the experiments on data rate we conclude that, in this case, there is little difference in the accuracy of speed estimation for a high rate vs. low rate of data points over the duration as can be seen in Figure 4. The fluctuations in both figures are likely attributed to the scarcity of trajectory data compared to data from Trafikverket. For certain roads, there were instances where almost no data was available for one or two hours between 01:00 and 03:00. This scarcity of data makes the mean speed susceptible to the influence of exceptionally fast or slow drivers during those periods, leading to fluctuations in the GPS data.

Despite having a statistically significant result from the $t$-test, the underlying assumptions making this test appropriate for this data need to be considered. There is a large difference in the variances of the data rates between the two groups tested, violating the assumption of the homogeneity of variances. Furthermore, the group with a mean velocity of less than 120km/h was not normally distributed and therefore violates the assumption of normality, which is essential for a two-sample $t$-test. Furthermore, to handle the pedestrian trajectories, the data points with a mean velocity $< 10$ km/h were excluded from the analysis. In doing so, motor traffic in congested states might also have been excluded, leading to non-representative results.

Were we to do this testing and analysis again, changes would have to be made. Instead of separating the trajectories into two groups by mean velocities, with the assumption that mean velocities $\geq 120$ km/h contain more erroneous data. Instead, after removing trajectories with a mean velocity $\geq 120$ km/h, we would have divided all data on the median data rate to have two equally large groups of assumed valid trajectories. Then perform the $t$-test on the RMSE of the two groups' mean velocity as compared to the ground truth from Trafikverket.

## 7.2 Hypothesis 2

We can estimate the annual average daily traffic with a RMSE of roughly 7442 cars. This mean error is much too large to be useful in most aspects of planning infrastructure or traffic. Most annual average daily traffic values have an error margin between 6-8%, so on the largest roads the mean error becomes acceptable and one could use a more current dataset to quite quickly get a very rough estimate of the current annual average daily traffic. On the other hand, one could use this kind of dataset to get insights into changes to current traffic patterns since the correlation between the counts is strong. An increase in volume in the dataset indicates an increase in real traffic. A relatively large count in the dataset indicates a relatively large annual average daily traffic.

To increase the accuracy of the estimation one should explore more complex models that take more features into account, both of the trajectories and the road segments, and of course train on many more road segments. This exploration began by extracting road networks and features from OpenStreetMap contributors (2023), but this increased the complexity of the task greatly, and manually extracting annual average daily traffic-training data is quite time-consuming so the exploration was abandoned in the scope of this project. In a larger project, one could get access to Trafikverket's data API and perhaps automate the extraction of training data.

## 7.3 Hypothesis 3

As in hypothesis 1, the deviation is present throughout the day compared to the true average speed. This is most likely because the GPS data is very sparse in comparison to the true data which is clear from the bar charts in Figure 6. Regarding the difference in speed drops, at a very late stage of this project, it was proposed by our supervisor that the reason for the major shifts in average speeds occurring about one hour before the validation data could be an error in the time zones. The main theory before this was that the difference was due to daylight saving time, but it is possible that the dataset we received had not been formatted to the Swedish timezone, GMT+1, and therefore had similar traffic flow patterns but not at the expected hour. As of writing this report, we have not received any further information on how the time zones of the original dataset were formatted.

The data from Trafikverket often goes above the speed limit of the roads. During the selection process for roads to serve as benchmarks, it was common to encounter roads with average speeds exceeding the set speed limit by 5-10 km/h. Consequently, the actual average velocities could have been different even if other days been selected. However, the selected days exhibited relatively stable speeds throughout, as evident when comparing the plots in Figure 6.

In a broader sense, the GPS data aligns somewhat well with the actual speeds. Especially for congested afternoon roads, a major speed reduction coincides with the morning rush hour timings, but is less present during afternoon rush hours when it is, in fact, greater based on the validation data. From these observations, it is reasonable to infer that this kind of data can indicate general trends in traffic flow states on freeways but is less reliable in determining the actual mean speed. It should be noted though, that speed is not always a direct indicator of traffic volume. Numerous external factors can lead to reduced speeds, irrespective of vehicle count, such as adverse winter conditions leading to slippery roads or other inclement weather scenarios.

## 7.4 General discussion

The inherently noisy nature of phone location data requires researchers to spend a lot of time on cleaning and pre-processing the data to get accurate distance- and speed metrics on trajectories. The complexity of this task was reduced in this project by implementing a very crude and simple map-matching technique based only on geographical closeness to the road section and the compass heading of the trajectory segments that intersect with the road section. This is most likely good enough for an exploratory analysis on larger and more isolated road sections as in this project, but on a road in the middle of the city one would also have to differentiate pedestrians and cyclists and as such would require a much more involved method of map-matching. A fourth hypothesis that stated that one should be able to do this differentiation without proper map-matching was initially explored in this project but later abandoned.

One glaring issue in this project is that the ground truth data does not match the dataset temporally. In most cases, this is not an issue since patterns do not seem to change much from year to year, but in some cases, we might have caught a road maintenance or construction in 2019 that is compared to data recorded in 2017 for example. A lot has happened recently in Gothenburg's road network (*e.g.* Västlänken, Hisingsbron, Marieholmstunneln) that have impacted traffic paths, flow, speed, and congestion patterns in different areas during different time periods before and after the dataset was recorded. To properly test this kind of dataset, one could probably coordinate with for example Trafikverket so that they do, or so that the researcher can access, actual measurements on some roads as ground truth that temporally match the dataset.

Also, experimental results by Tu et al. (2021) indicate that estimates begin to approach

accurate values after a GPS penetration rate of 2-3%. A rough estimate of the GPS penetration rate of our dataset on the actual traffic population gives an average rate of about 0.03% which is clearly below this rate.

# 8 Conclusion

The main purpose of this project was to investigate if sparse GPS data is of good enough quality to estimate traffic flow states. This was divided into three sub-tasks, to see if a high ratio of data points over duration yields average velocities closer to the ground truth than a low ratio of data points over duration, to see if the dataset is good enough to estimate annual average daily traffic, and if it is possible to model hourly traffic flow.

The results indicate that (1) higher data rate trajectories seem to give estimations slightly closer to ground truth but the difference in RMSE has not been statistically tested, (2) annual average daily traffic can be estimated with a large mean error and the correlation between the dataset and ground truth is strong, and (3) hourly speed- and flow traffic patterns can be roughly captured.

Further research of interest could focus on the differentiation of pedestrians, cyclists, and motor traffic with the implementation of more proper map-matching than the map-matching done in this study. The motivation for this is to estimate and analyze traffic flow on city roads that were excluded in this report. Also, due to the temporal mismatch between the ground-truth data from Trafikverket and the data used in this report, coordinating with Trafikverket might provide actual measurements more accurate to the real-world setting to further validate the dataset.

# Acknowledgements

This project would not have been possible without our advisor, Jorge Gil, who provided generous feedback, expertise, and support. Map data copyrighted OpenStreetMap contributors and available from https://www.openstreetmap.org.

# References


Chao, P., Xu, Y., Hua, W., & Zhou, X. (2020). A Survey on Map-Matching Algorithms. In R. Borovica-Gajic, J. Qi, & W. Wang (Eds.) *Databases Theory and Applications*, (pp. 121–133). Cham: Springer International Publishing.

Coppola, R., & Morisio, M. (2016). Connected car: Technologies, issues, future trends. *ACM Computing Surveys*, *49*(3).

den Bossche, J. V., Jordahl, K., Fleischmann, M., McBride, J., Wasserman, J., Richards, M., Badaracco, A. G., Snow, A. D., Tratner, J., Gerard, J., Ward, B., Perry, M., Farmer, C., Hjelle, G. A., Taves, M., ter Hoeven, E., Cochran, M., rraymondgh, Gillies, S., Caria, G., Culbertson, L., Bell, R., Bartos, M., Eubank, N., sangarshanan, Flavin, J., Rey, S., Gardiner, J., maxalbert, & Bilogur, A. (2023). geopandas/geopandas: v0.14.0.
URL `https://doi.org/10.5281/zenodo.8348034`

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Second International Conference on Knowledge Discovery and Data Mining (KDD'96). Proceedings of a conference held August 2-4*, (pp. 226–331).

Gillies, S., van der Wel, C., den Bossche, J., Taves, M. W., Arnott, J., Ward, B. C., & others (2023). Shapely.
URL `https://github.com/shapely/shapely`

Gkountouna, O., Pfoser, D., & Zufle, A. (2020). Traffic flow estimation using probe vehicle data. In *Proceedings - 2020 IEEE 7th International Conference on Data Science and Advanced Analytics, DSAA 2020*, (pp. 579–588). Institute of Electrical and Electronics Engineers Inc.

Greenshields, B. D., Thompson, J. T., Dickinson, H. C., & Swinton, R. S. (1934). The Photographic Method Of Studying Traffic Behavior. In *Highway Research Board Proceedings*, vol. 13.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95.

Jiang, L., Chen, C.-X., & Chen, C. (2023). L2MM: Learning to Map Matching with Deep Models for Low-Quality GPS Trajectory Data. *ACM Trans. Knowl. Discov. Data*, *17*(3).
URL `https://doi-org.ezproxy.ub.gu.se/10.1145/3550486`

Kerner, B. S., Klenov, S. L., & Hiller, A. (2007). Empirical test of a microscopic three-phase traffic theory. *Nonlinear Dynamics*, *49*(4), 525–553.

Kessels, F. (2019). The Fundamental Diagram. In *Traffic Flow Modelling: Introduction to Traffic Flow Theory Through a Genealogy of Models*, (pp. 21–34). Cham: Springer International Publishing.
URL `https://doi.org/10.1007/978-3-319-78695-7_2`

Kühne, R. D. (2011). Greenshields' Legacy: Highway Traffic. In G. J. Beal (Ed.) *75 Years of the Fundamental Diagram for Traffic Flow Theory. Greenshields Symposium.*, (pp. 3–10). Woods Hole, Massachusetts.

Murphy Kevin P. (2012). *Machine Learning A Probabilistic Perspective* . Cambridge: The MIT Press.

OpenStreetMap contributors (2023). Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org.

Ozdemir, S., Kakade, S., & Tibaldeschi, M. (2018). *Principles of Data Science*. Birmingham: Packt Publishing Ltd., second ed.

QGIS.org (2023). Qgis geographic information system.
URL https://www.qgis.org

Sharma, S., & Kaushik, B. (2019). A survey on internet of vehicles: Applications, security issues & solutions. *Vehicular Communications*, *20*.
URL https://doi.org/10.1016/j.vehcom.2019.100182

Skiena, S. S. (2017). *The Data Science Design Manual*. Cham: Springer International Publishing.

Trafikverket (2015). Metodbeskrivning - undersökningen av ÅDT. Tech. rep., Trafikverket.
URL https://bransch.trafikverket.se/contentassets/29f030c5f81948f1ae48ab3b73c02d30/metodbeskrivning_adt_2015-06-18.pdf

Trafikverket (2022). About Us.
URL https://bransch.trafikverket.se/en/startpage/about-us/Trafikverket/

Trafikverket (2023). Vägtrafikflödeskartan.
URL https://vtf.trafikverket.se/SeTrafikinformation

Tu, W., Xiao, F., Li, L., & Fu, L. (2021). Estimating traffic flow states with smart phone sensor data. *Transportation Research Part C: Emerging Technologies*, *126*, 103062.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: Createspace.

Wang, Z., Lu, M., Yuan, X., Zhang, J., & Wetering, H. V. D. (2013). Visual traffic jam analysis based on trajectory data. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12).

Xing, J., Liu, Z., Wu, C., & Chen, S. (2019). Traffic Volume Estimation in Multimodal Urban Networks Using Cell Phone Location Data. *IEEE Intelligent Transportation Systems Magazine*, *11*(3), 93–104.

Yuan, H., & Li, G. (2021). A Survey of Traffic Prediction: from Spatio-Temporal Data to Intelligent Transportation. *Data Science and Engineering*, *6*(1), 63–85.
URL https://link.springer.com/article/10.1007/s41019-020-00151-z

Zhao, L., Ochieng, W. Y., Quddus, M. A., & Noland, R. B. (2003). An Extended Kalman Filter Algorithm for Integrating GPS and Low Cost Dead Reckoning System Data for Vehicle Performance and Emissions Monitoring. *Journal of Navigation*, *56*(2), 257–275.

Zheng, Y. (2015). Trajectory Data Mining. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *6*(3).
URL https://dl-acm-org.ezproxy.ub.gu.se/doi/10.1145/2743025

# A Map of measured road sections

**Figure 7**
*Selected road sections plotted on a map of Gothenburg. Red circles are congested, blue circles are free-flow.*