Data-driven traffic flow: Summary of experiments

TENS-PROJECT

Ian Marsh, Henrik Abrahamsson (RISE AB)
Pei-Lun Hsu, (KTH)

RISE report 2021:119
ISBN 978-91-89561-10-6

English abstract

This final report is the *second* of two reports. Both are the result of a Swedish project, the Swedish Energy Authority's TENS project, which spanned from 2018-2021. Other reports from the project provide results from estimating emissions from traffic measurements as well as simulation studies. Like any European capital, Stockholm suffers from many problems related to its road network. The main factor is traffic jams, which are aggravated with difficult weather conditions in winter but also due to accidents, popular events and holidays. Therefore, this report provides the results from a data-driven approach to estimating traffic flow. This work aims at predicting and understanding the behavior of this network based on data collected at several places. More specifically, the goal is to predict and model the traffic flow i.e macroscopic information, on ground measurements (MCS), using floating microscopic (INRIX) data. We focus on estimating the fundamental traffic law relationships, the flow using time series and future directions. Methods and results are in the related work section.

Keywords: Transport ITS, data-driven, traffic flow.

Svenska sammanfattning

Denna slutrapport är den andra av *två* rapporter. Båda är resultatet av två svenska projekt, Energimyndighetens TENS-projekt, som sträckte sig från 2018-2021. Andra rapporter från projektet ger resultat från uppskattning av utsläpp från trafikmätningar och simulering. Som vilken europeisk huvudstad som helst, lider Stockholm av många problem relaterade till vägnätet. Den främsta är trafikstockningar, som förvärras med svåra väderförhållanden till exempel vintertid, men också när olyckor inträffar. Därför ger den här rapporten resultaten från en datadriven metod för att uppskatta trafikflödet. Detta arbete syftar till att förutsäga och förstå beteendet i detta nätverk baserat på data som samlats in på flera ställen, med hänsyn till de olika variabler som påverkar hur människor kör. Mer specifikt är målet att förutsäga och modellera trafikflödet, dvs makroskopisk information, på markmätningar (MCS), med hjälp av flytande mikroskopisk (INRIX) data. Vi fokuserar på att uppskatta de grundläggande trafiklagsförhållandena, flödet med tidsserier och framtida riktningar. Metoder och resultat finns i det relaterade arbetsavsnittet.

Nyckelord: Transport ITS, datadriven, trafikflöde.

The TENS project
Energimyndigheten
Diarienummer: 2018-006615
Projektnummer: 46963-1

1 Introduction

The goal with the analysis is to observe traffic in practice and how flow relates to speed. First looking at the MCS flow and MCS speed measurements in order to:

1. Predict flow from speed (without the extra complications and uncertainties that are introduced when trying to use INRIX speed to estimate MCS flow)
2. Can speed variance or time of day be useful features to improve the flow prediction?
3. Test different methods for prediction

Traffic flow on a motorway is often very predictable from day to day. There is little traffic during night, and there are rush hours in the mornings and afternoons during weekdays. The pattern is different during weekends and holidays with less traffic in the mornings. The average speed on the road follows a similar daily cycle. There are high speeds around the speed limit during night and during other free-flow periods, and low speeds when there is congestion during the rush hours. The figures below show how the traffic flow and average speed varies over one week on a motorway section in Stockholm in October 2018.

2 The datasets

MCS data

The traffic flow in Stockholm is monitored with a Motorway Control System (MCS). A large number of stationary MCS-portals gantries have been installed on the E4 and other roads. The MCS-portals are equipped with radar detectors and they monitor the flow and speed of traffic in each lane of the road. The data gives the regional traffic control centre information about the current traffic flow and speeds. The data is also input to a control system that sets variable speed limit signs. The radar measurements are point measurements and give time-mean-speed. This is in contrast to probe measurements such as INRIX calculate speed over a distance and provide space-mean speed. The Stockholm Motorway Control System, including the subsystems for Variable Speed Limits (VSL) and Automatic Incident Detection (AID), is described in [Nissan2010].

The Motorway Control System (MCS)

ID	Date	Speed	Speed Deviation	Flow	Used Lanes
1164	2018-10-01 00:01:00	16.94	7.18	120	{1}
1164	2018-10-01 00:22:00	13.05	9.53	60	{1}

Table 1: An example of the MCS dataset.

INRIX data

Founded in 2005, INRIX pioneered the practice of managing traffic by analyzing data not just from road sensors, but also from vehicles. This breakthrough approach enabled INRIX to become one of the leading providers of data and insight into how people move around the world. INRIX combines data from many sources to provide traffic information. A major part of the data comes from a crowdsourced model where INRIX continuously collects speed and location from probe vehicles, combines the data into an updated view of the current traffic situation on the road, and sends it back to the vehicles.

The INRIX data

Segment ID	TMCID Timestamp (UTC)	Segment Type	Speed	Average	Referenced speed	Travel time (mins)	Score	Cvalue	Speed Bucket	Registered
225525816	2018-10-0 1 00:00:15	XDS	85	0.501	60	30	56	49	3	2018-10-0 1 02:00:14
225525816	2018-10-0 1 00:01:13	XDS	95	0.501	68	30	56	98	3	2018-11-0 1 00:55:11

Table 2: An example of the INRIX dataset.

Field Descriptions

Example Segment ID An identifier for defining a unique road segment. Timestamp The timestamp of the measurement in UTC. Segment Type Type of road segment: XD Segment (XDS) or TMC segment. XDS Speed Average speed of vehicles on the segment calculated from the most current time slice, in km/h. Average Speed The historical average speed on the segment for the given day and time (km/h). Reference Speed An expected free-flow speed on the segment, determined from the INRIX traffic archive (km/h). Travel Time. The time required to travel across the segment in minutes. Score. A measure of confidence in a given reported speed with three possible values: 10/20/30. Samples with confidence scores larger than 10 are based on real-time data, otherwise are based on historical data. Cvalue. The second measure of confidence ranges from 0 to 100, which only applies when the confidence score is 30. Speed-bucket Level of congestion according to the range of speed. Level of congestion according to the range of speed.

MCS and INRIX locations

The goal of this report is to evaluate using MCS and INRIX together. That is using INRIX speed and flow to infer MCS speed and flow.

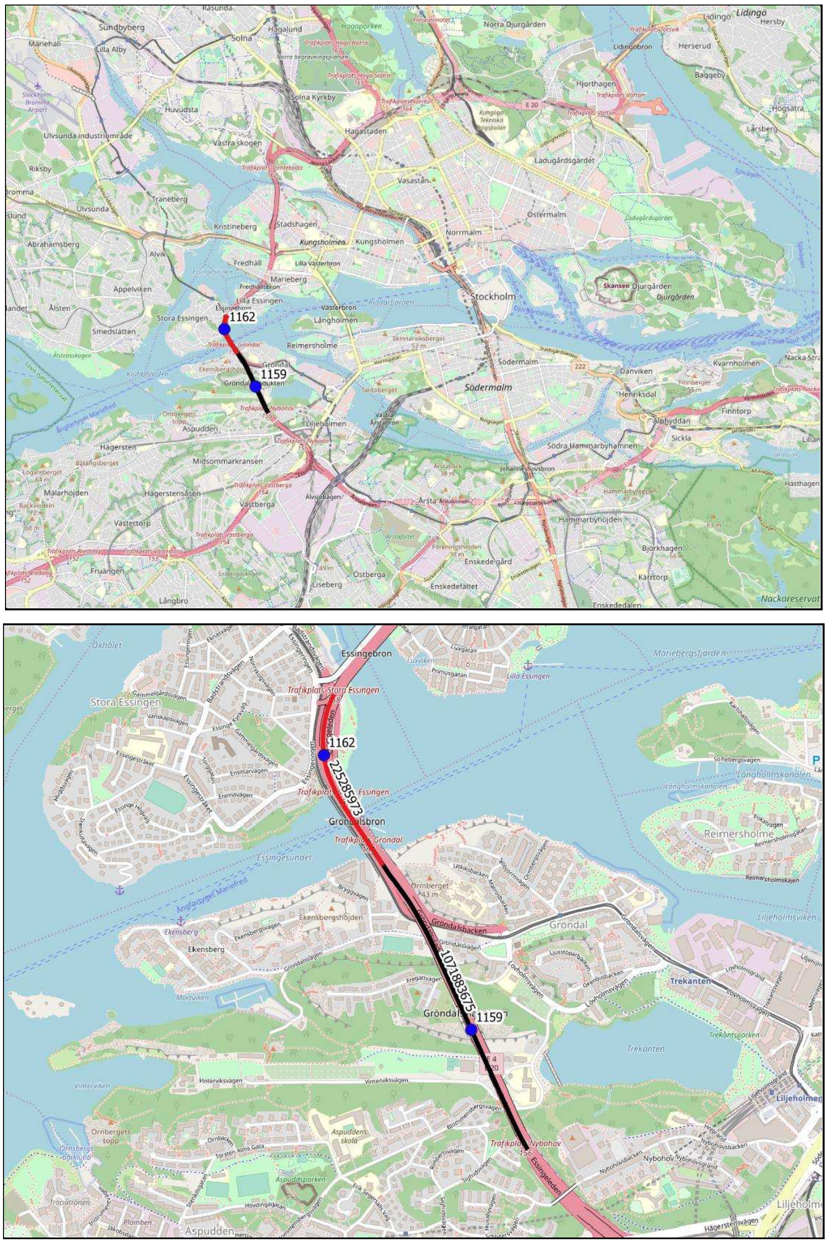

Figure 1: Locations of the MCS and INRIX data segments in Stockholm

3 Traffic state estimation

Figure 2 is a conceptual diagram which describes the traffic data, traffic state, and their relation in the time-space domain. As mentioned before, a subset of traffic state variables can represent a traffic station highway, i.e., any two or three traffic state variables. In figure 2, the traffic state is observable at the locations where stationary sensors are installed. However, stationary sensors are usually installed sparsely on highways due to cost reasons. In the regions outside stationary sensors' coverage, the yellow regions in figure 2, traffic state variables are either unobserved or only partially observed through traffic data collected by other sensors, e.g., mobile devices. TSE aims to estimate the unobserved traffic state variables using partially observed traffic data in regions where stationary sensors are absent, e.g., the yellow region with the dashed boundary in the figure. Moreover, even at the locations where stationary sensors are installed, traffic datasets usually suffer from missing or corrupted data due to failures of sensors or communication. Therefore, a specific TSE method called imputation has been developed for imputing the missing or corrupted data in the traffic datasets collected from the locations where the sensors are installed but suffer from missing data.

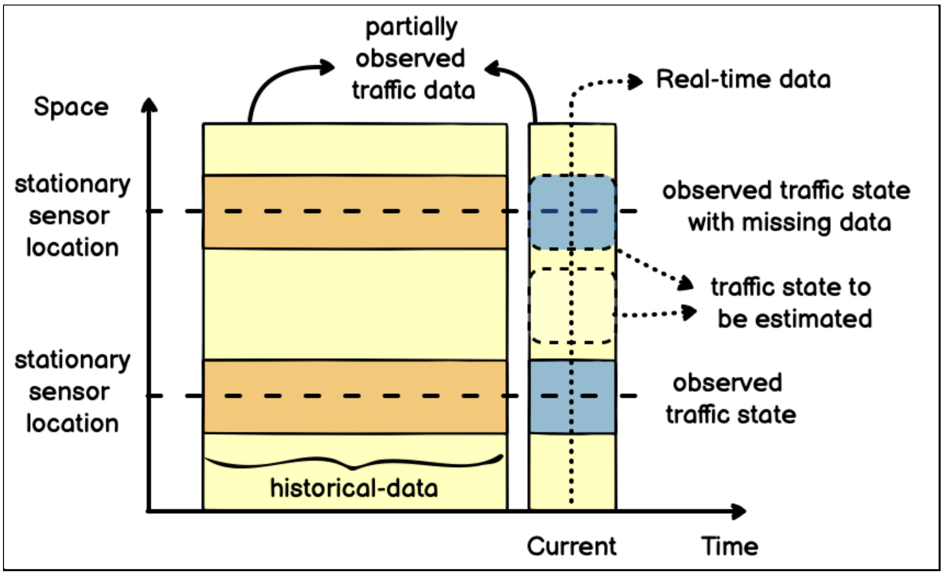

Figure 2: An illustration of the traffic state estimation (adapted from [Seo2017]).

4 Data exploration

We first look at the MCS and INRIX data from an exploratory point of view. The point of views are times series, time correlations, speed histograms.

Time series view

If we look at the first 21 days in the dataset, we see that the traffic is very similar from week to week on an hourly basis.

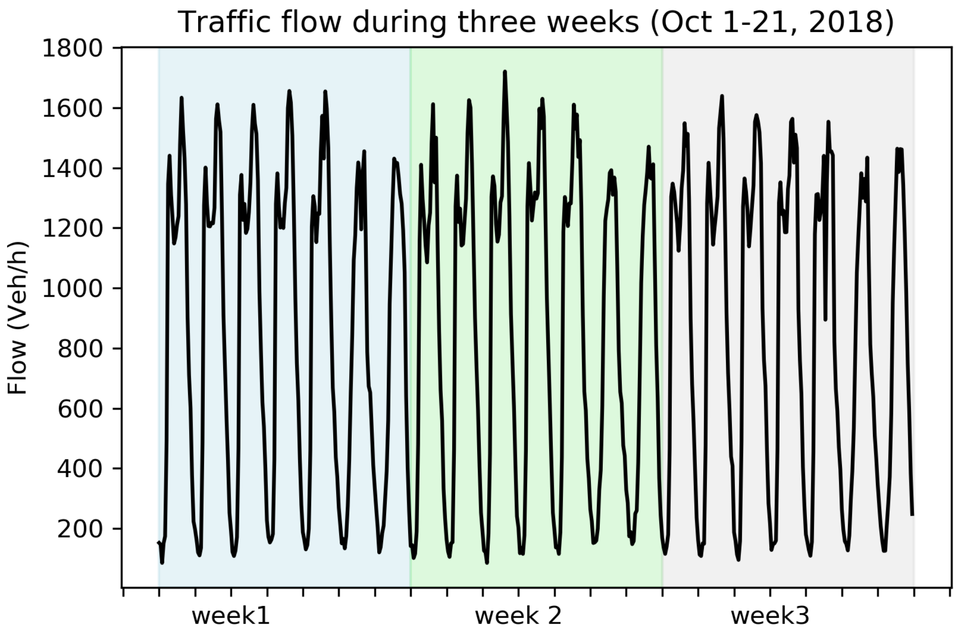

Figure 3: Traffic flow over three weeks in 2018

In the graph below is the data by weekday and calculated the average flow per hour, based on the first three weeks of data (Oct 1-21). We have the average flow on Mondays between 00:00-01:00, 01:00-02:00 etc., for each hour per week. We can see that the traffic behavior on weekdays Monday-Thursday is very similar. Friday is somewhat different, indicating an earlier second rush hour. And weekends are different with much less traffic in the mornings.

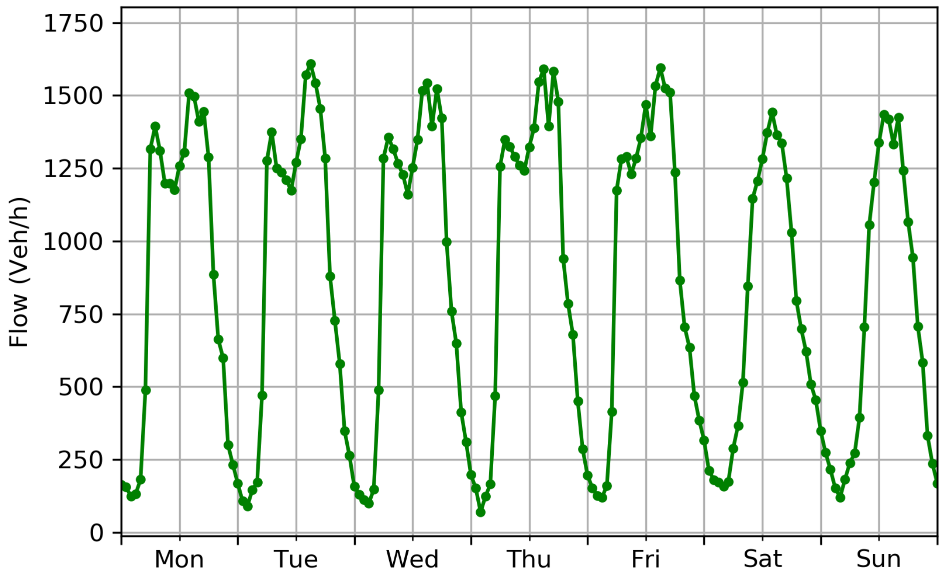

Figure 4: Flow of vehicles over one week in 2018

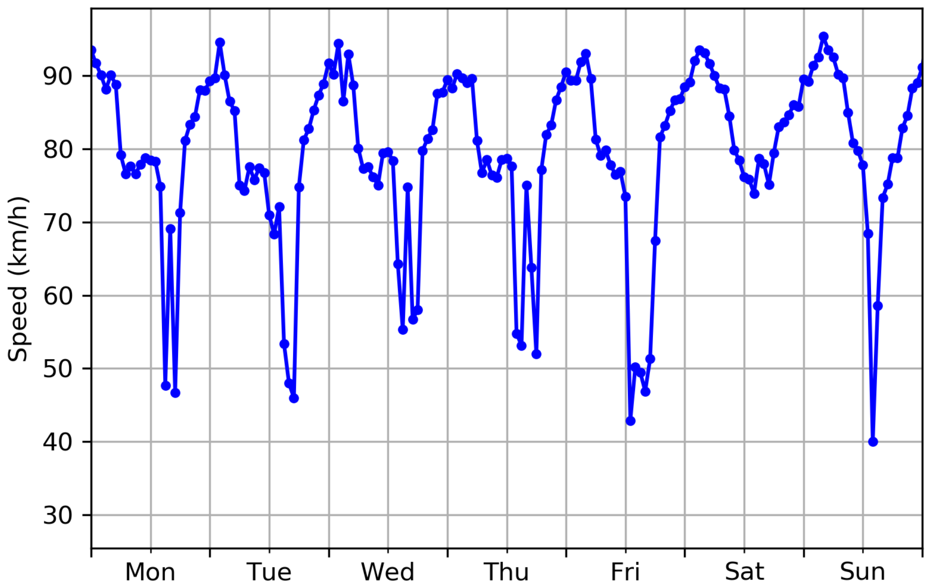

Figure 5: Speed of vehicles over one week in 2018

The daily and weekly traffic patterns can of course vary between different motorways. Not all motorways have congestion during rush hours. Some roads have lower speeds during night because there is a large share of slower trucks and heavy traffic on the road at that time, etc. The amount of traffic and the average speed may also depend on the weather and the road conditions. For instance, with lower speeds during winter when there is snow and slippery roads. The traffic pattern on a motorway with daily and weekly cycles is very predictable. But it can also change both temporarily and over a long period of time. It is therefore important to continuously monitor the traffic.

The traffic can change on an hourly basis due to accidents or sport events or weather conditions. It can change for weeks and months due to roadworks or as a more unusual example: due to a pandemic[1]. There are also many long-term trends and political decisions that might influence the traffic flow. Just a few examples: more people are moving to cities which means more traffic & subsequently congestion. New residential areas are built but also alternative roads, that in some cases move traffic away from cities. There is a trend to cycle or use public transport instead of using vehicles, for obvious environmental reasons. There are political decisions such as road tolls, bridges & tunnel projects that influence traffic patterns within a city.

Time corrections

As we can see the timestamps, over and above the time zone differences diverge. We will later compensate for it too. A simple subtraction can be done, but as one sees in the Figure, the steps indicate perhaps more advanced methods such as Dynamic Time Warping could be applied [*DTW*].

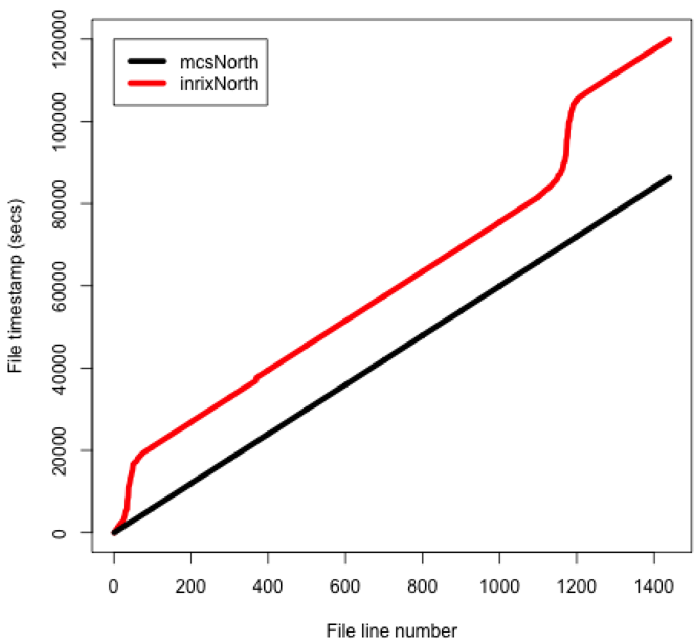

Figure 6: Timestamp misalignment example of the MCS and INRIX datasets

[1] The traffic on the Swedish road network decreased with 25% in April 2020 due to the Corona virus <https://www.trafikverket.se/tjanster/trafiktjanster/Vagtrafik--och-hastighetsdata/trafikarbetets-forandring-pa-det-statliga-vagnatet-tf/trafikforandringar-under-coronaviruset/ >, viewed 7 August 2020.

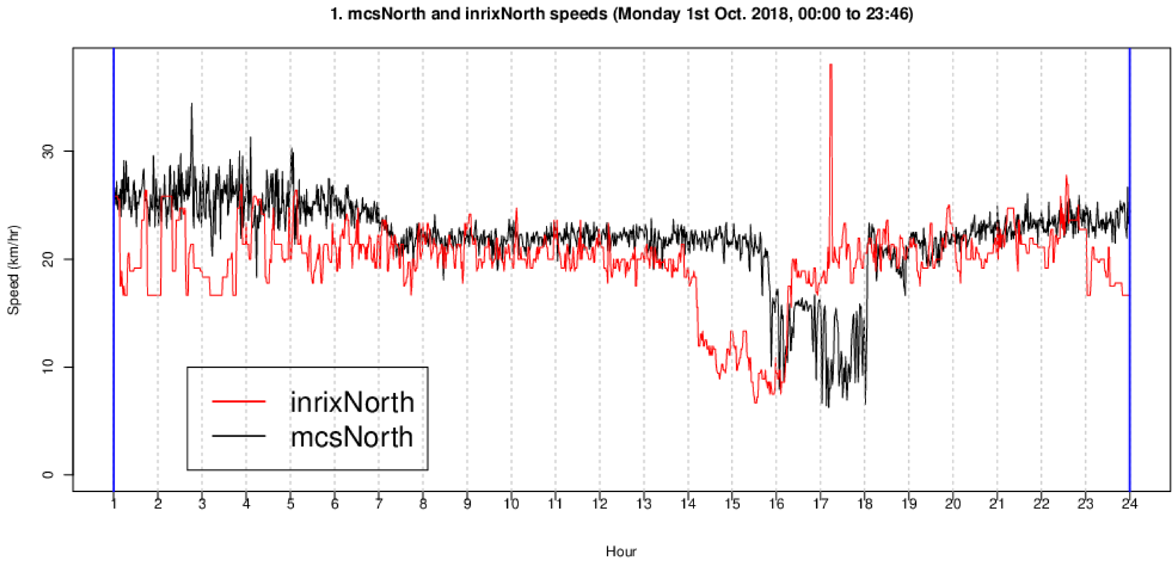

Figure 7: Timezone differences of the INRIX and MCS datasets

Time windowing

In part I of these reports, we posed the question what are the appropriate parameters with which to process the data. Windowing is important as the intervals range from 1 to 1440 minutes per day. The length of a window can have significant effects on the outcomes, whether on the averages, smoothing, input to the ML or the visualistions.

There is an extra complication. INRIX speed and flow are given over a segment whilst MCS speed and flow are from a point (in general). Both need some processing. For simplicity we will choose the same chunks at which we look at. In part I of the report we gave a glossary and their units and typical values. Some need inputs per minute, but are acted upon on 10-15 minute intervals. Obviously, speed control per minute would lead to road chaos. Note also, at least MCS reports speed and flow at resolutions less than a minute, we however only see that resolution.

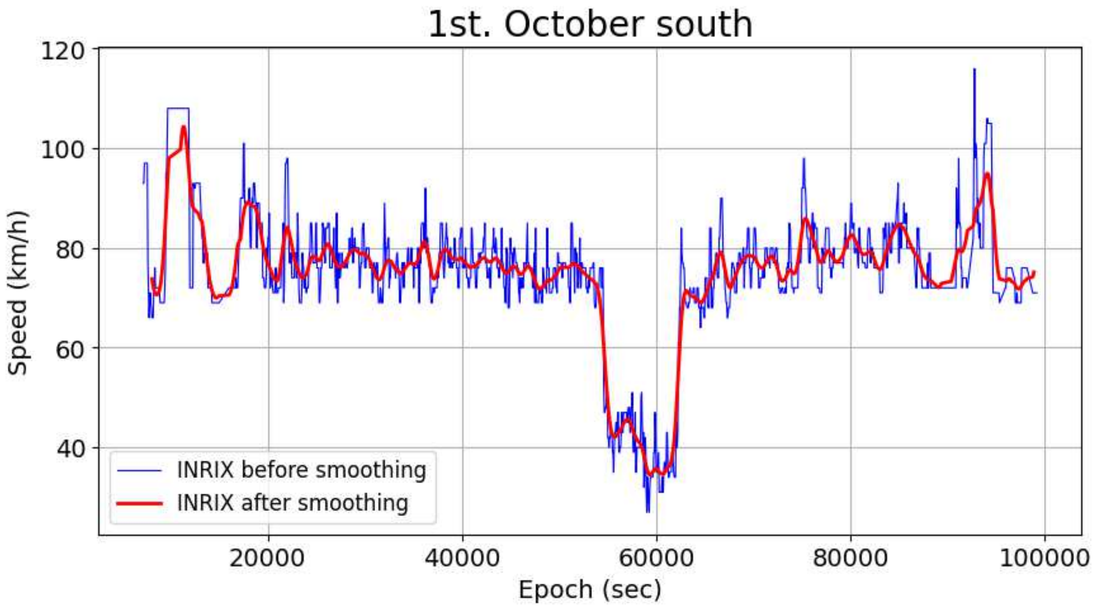

Figure 8: Example of smoothing the INRIX data

Time averaging/smoothing

The basic time resolution in traffic theory is the minute. Data collation, not collection, is often at this minute resolution. From each minute to hours, days to weeks some form of averaging or windowing is needed. Different time scales serve different purposes. Minute level-collations are processed hourly for rush-hour analysis. Daily aggregates are needed for weekday/weekend analysis etc. Averaging traffic flow quantities is needed. The time-mean speed is measured at a MCS reference point on the roadway over a period of time. Average speed measurements obtained from this method are not accurate because instantaneous speeds averaged over several vehicles do not account for the difference in travel time for the vehicles that are traveling at different speeds over the same distance. The space-mean speed is measured over the whole roadway segment. Some form of monitoring tracks the speed of individual vehicles, and then the average speed is calculated. It is considered more accurate than the time mean speed. The time mean speed is never less than space mean speed, see [*Marsh2021TENS1*], section 6.

Week day flow differences

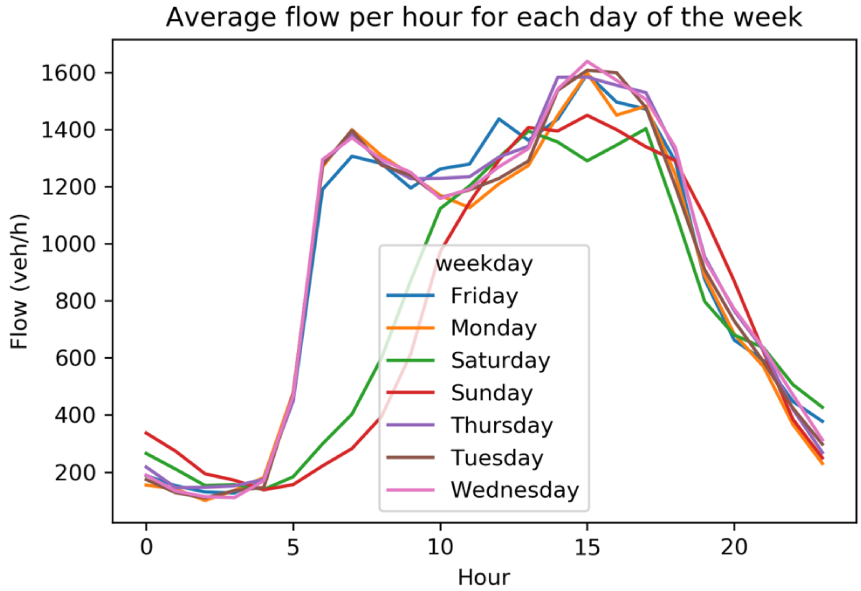

Figure 9: Flow over a day for each day

Speed differences between INRIX and MCS

Looking at the speed differences between the two types of flow we have, adjusted using the time shifting above we see the speed differences in Figure 7.

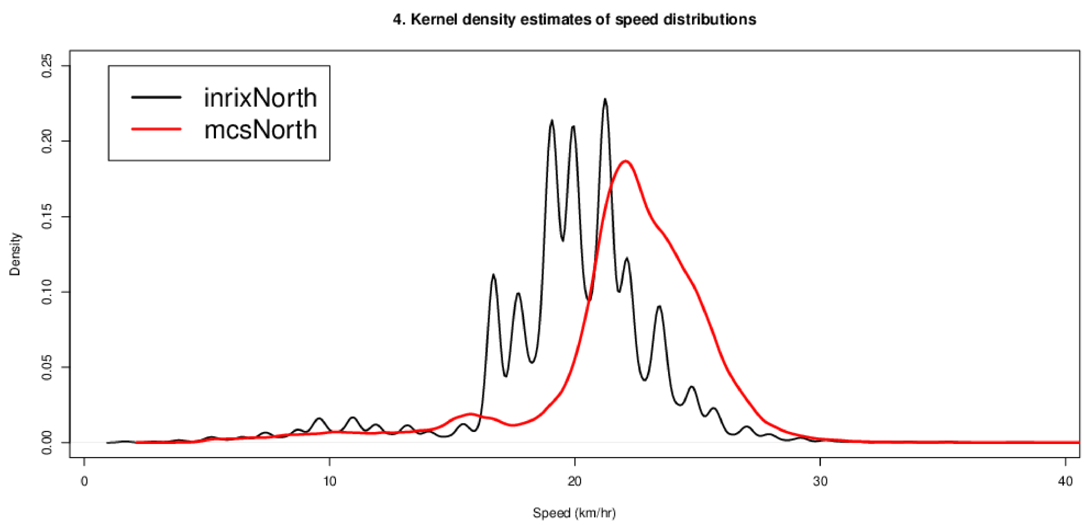

Figure 10: Example speed distributions of INRIX-MCS data using kernel densities

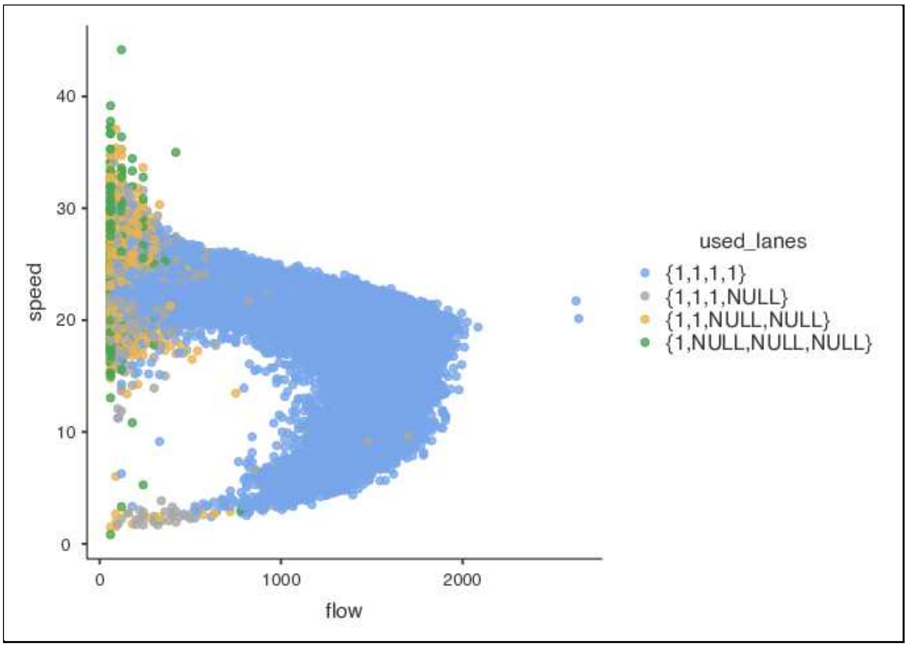

Figure 11: Flow versus speed (effect of 4, 3, 2 and 1 lanes)

Figure 11 shows how important it is to gather all data across the motorway. The blue points include all lanes including the slowest one. This gives the flow speed plot its distinctive shape seen in the theory texts. The texts however do not usually point out the lane importance in the theoretical plots.

5 Results: Estimating the dual phase traffic relationships

This section aims to introduce machine learning methods utilized in this work for extracting dynamic relations between traffic variables for traffic flow estimators. Below we introduce two models, i) piecewise linear regression, derived from the fundamental diagrams and ii) deep neural networks. Our works have looked at deep neural networks before, for example LSTM models to infer traffic flow. In this case for MCS data, alone [*Ghandeharioon2018Evaluation*].

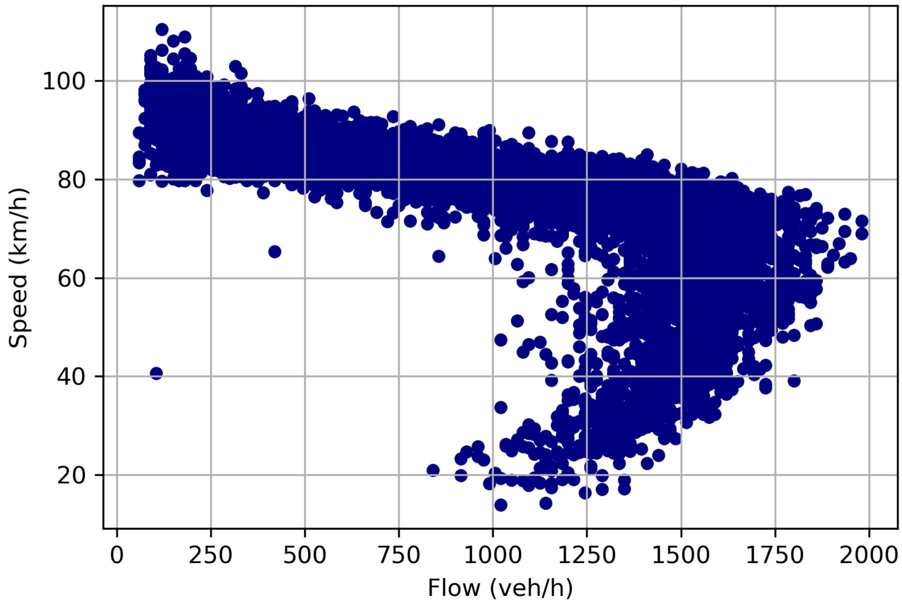

Figure 12: Flow versus speed (using the full dataset)

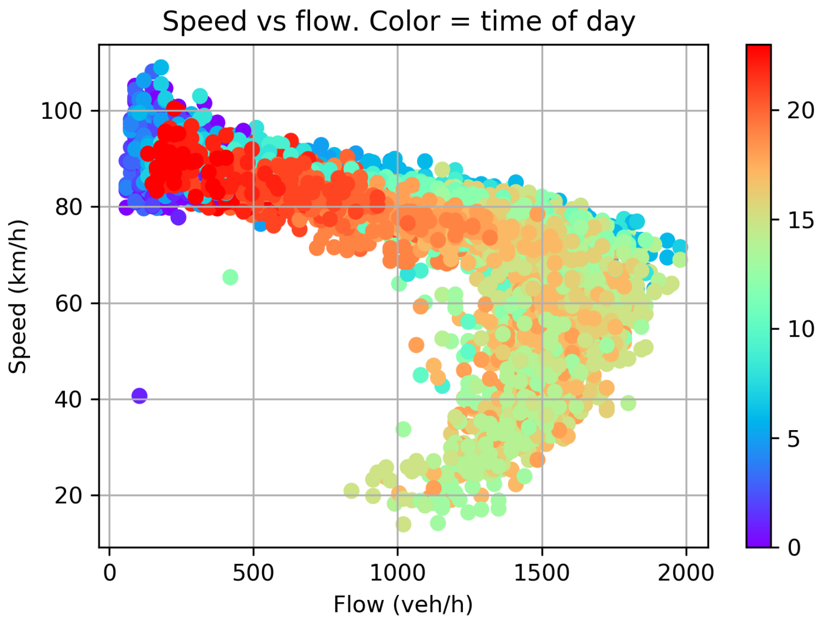

Figure 13 Flow versus speed (colour indicates time of day)

Estimating flow from speed using a two-step linear regression on the timescale of hours, re-sampled the MCS data to the timescale of hours. This resulted in 744 data points. used the first 80%, the first 25 days, as training data and the last 20% as test data.

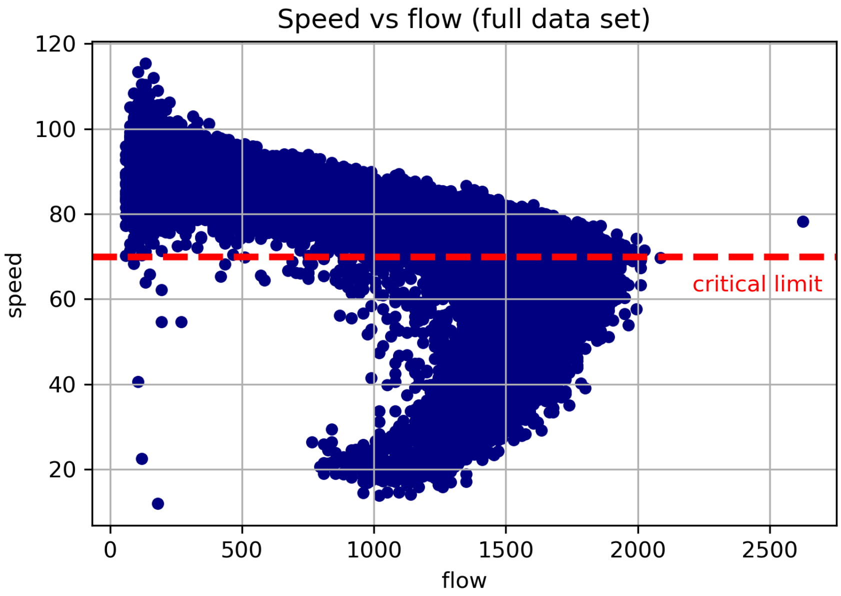

Figure 14: Flow versus speed (speed critical limit shown)

A first attempt to predict flow from speed using linear regression uses one line segment for the free flow state and another one for the congested state. A two-step linear regression model. Calculated one line for the free-flow phase, for data points where the speed was above the critical limit = 70km/h, and another line for the congested phase. With a linear model, high speeds could be mapped to negative flows. set the minimum predicted flow to 104 veh/h, which is the minimum flow value in the training data when re-sampled to hours.

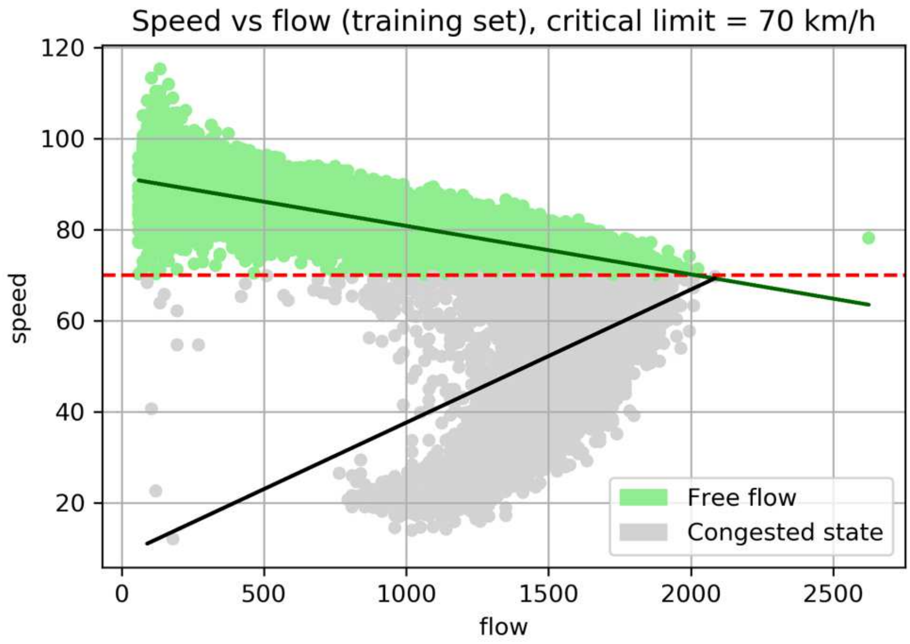

Figure 15: Flow versus speed (with critical limit at 70 km/h)

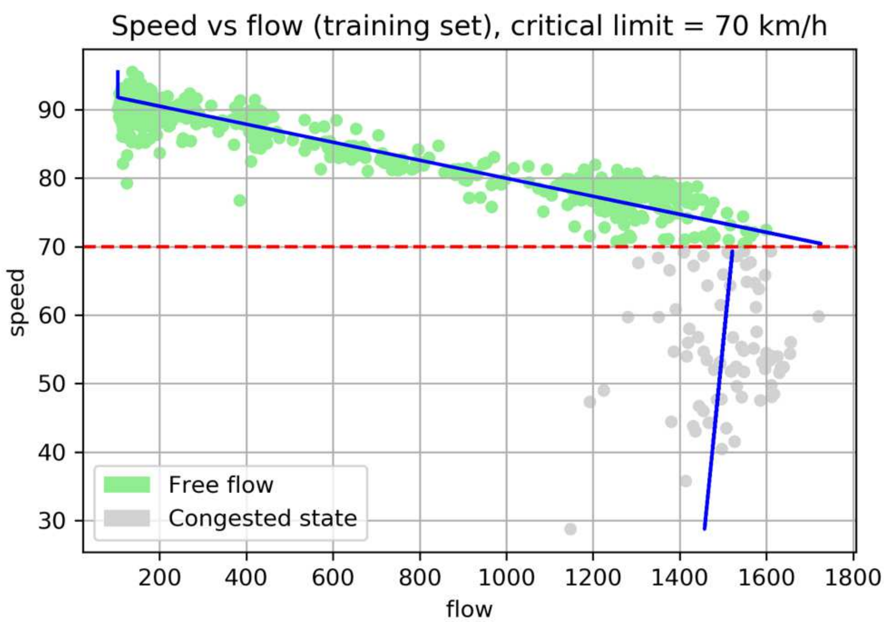

Figure 16 Flow versus speed (colour indicates time of day)

The same thing but plotting flow versus speed (since we predict flow from speed) :

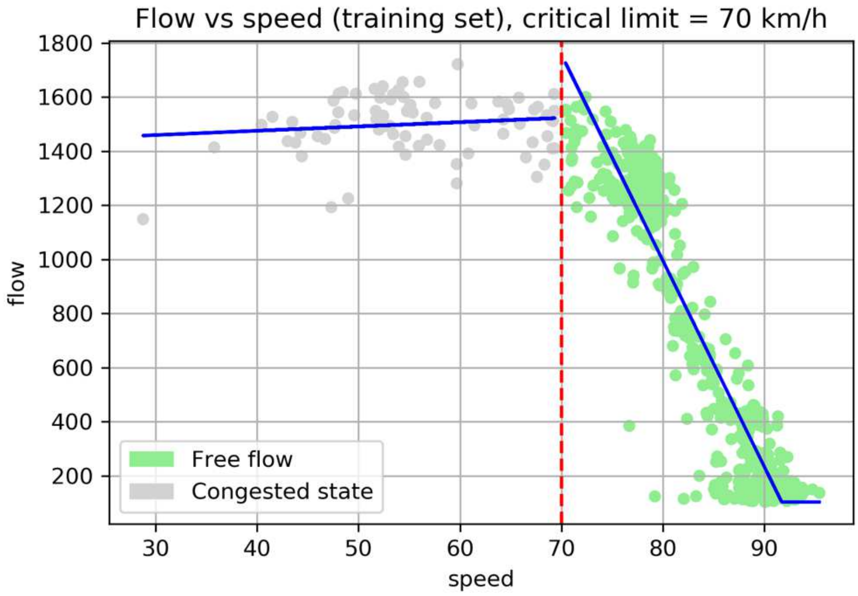

Figure 17 Flow versus speed (showing critical speed and flow indicated)

Evaluation of the two-step linear regression model

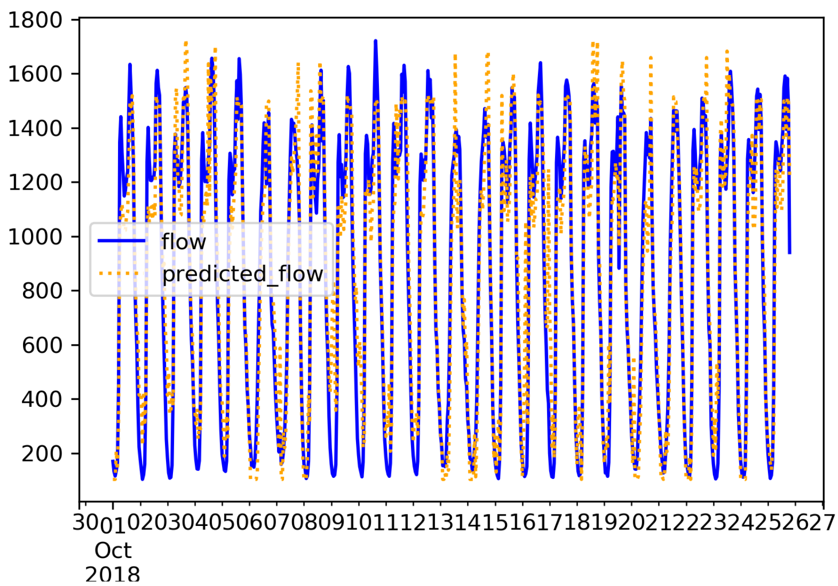

Figure 18 Flow prediction over a month (blue recorded flow, orange, a prediction)

Evaluation of the two-step linear regression model using the test set

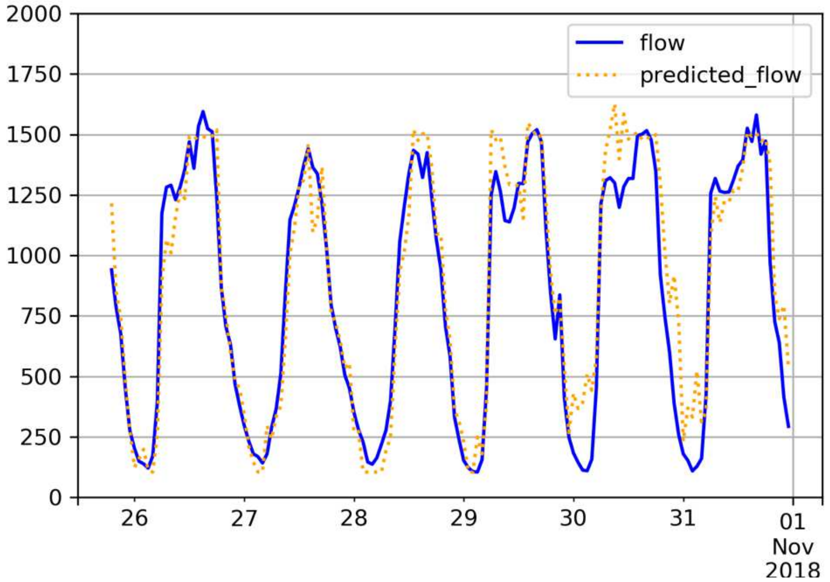

Figure 19 Flow prediction over a week, blue recorded flow, orange, a prediction, mean *absolute error: 114.96, R^2 score: 0.906*

Using speed variance to identify flow in the uncongested state

Blandin et. al. presents the idea that, in the uncongested phase, the drivers can drive at whatever speed they feel comfortable with, given the speed limit, if there is no or little other traffic [Blandin2012]. Hence the speed variance is expected to be high if the flow is low. However, at higher flows and still in the uncongested phase "*the individual speed variance is expected to be relatively small because commuters are constrained by other surrounding vehicles and hence cannot freely choose their traveling speeds.*"

In the MCS data we have the speed standard deviation as input for INRIX: so one calculates the variance when the data is upsampled to an hourly resolution. Looking at the speed standard deviation in the first week of the MCS data, 1st-7th of October 2018. Plotting the speed standard deviation over time and comparing it with speed and flow:

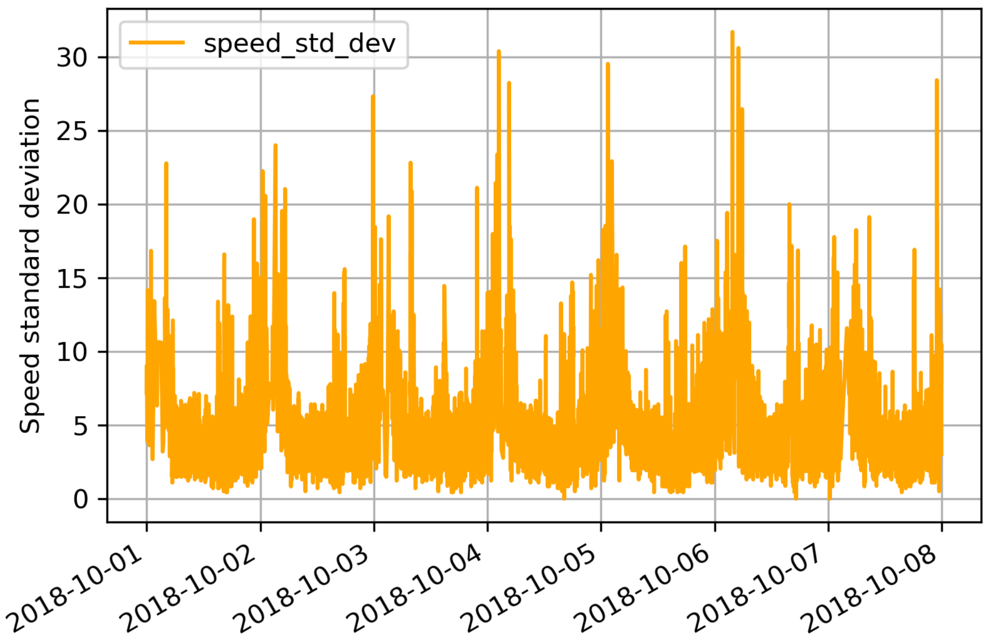

Figure 20 The speed standard deviation

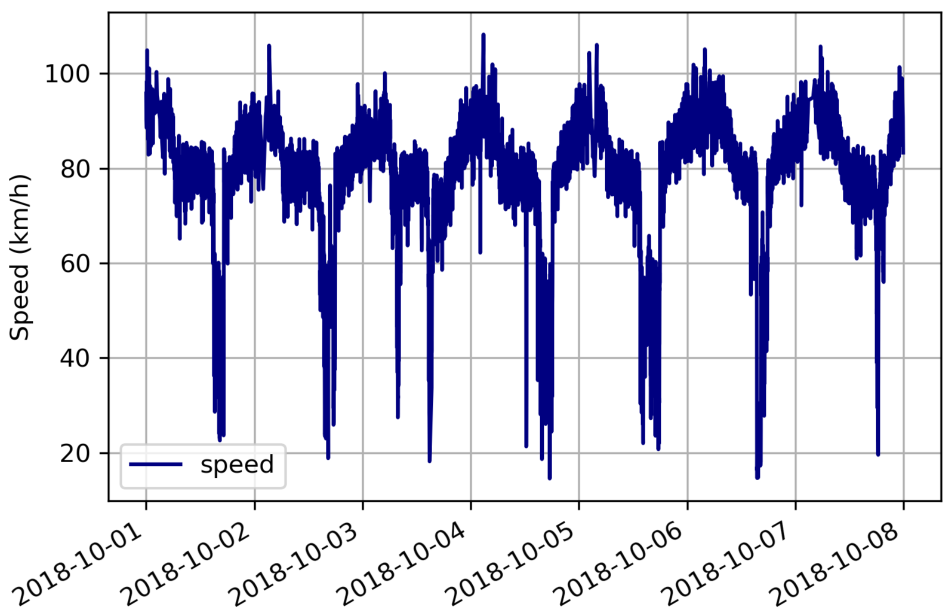

Figure 21 Speed variation over a week

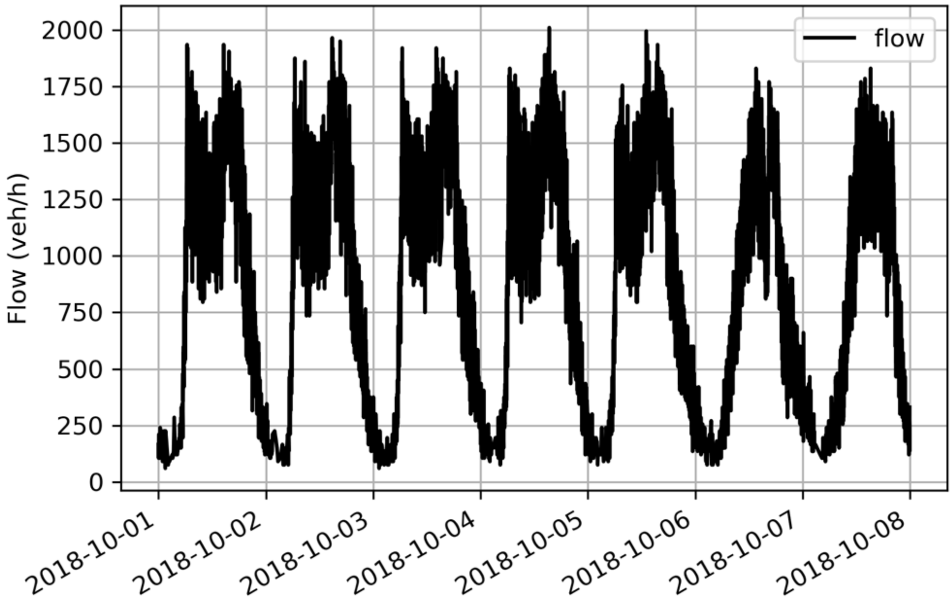

Figure 22 Flow variation over a week

One thing we notice is that the speed variance is higher during night when the speed is high. Below we see a scatter plot of speed versus flow. And the color shows the speed variance.

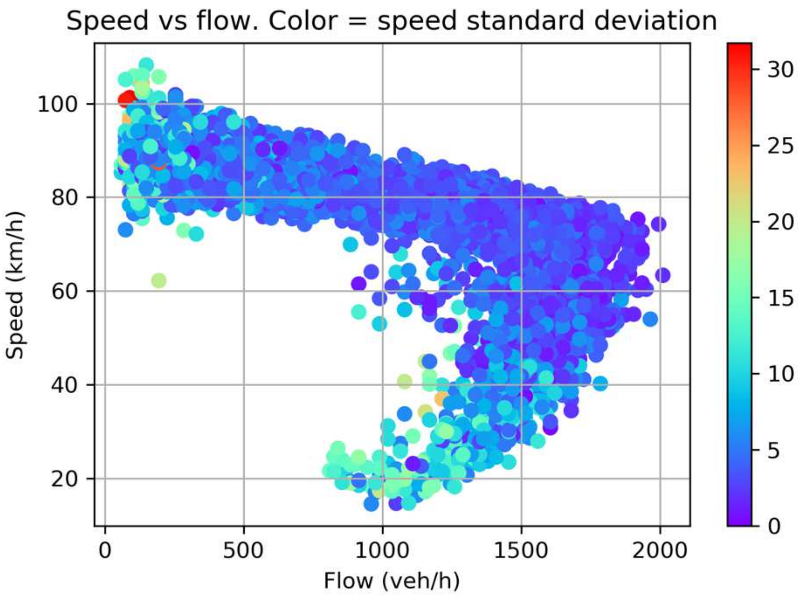

Figure 23 Flow versus speed with standard deviation as colours

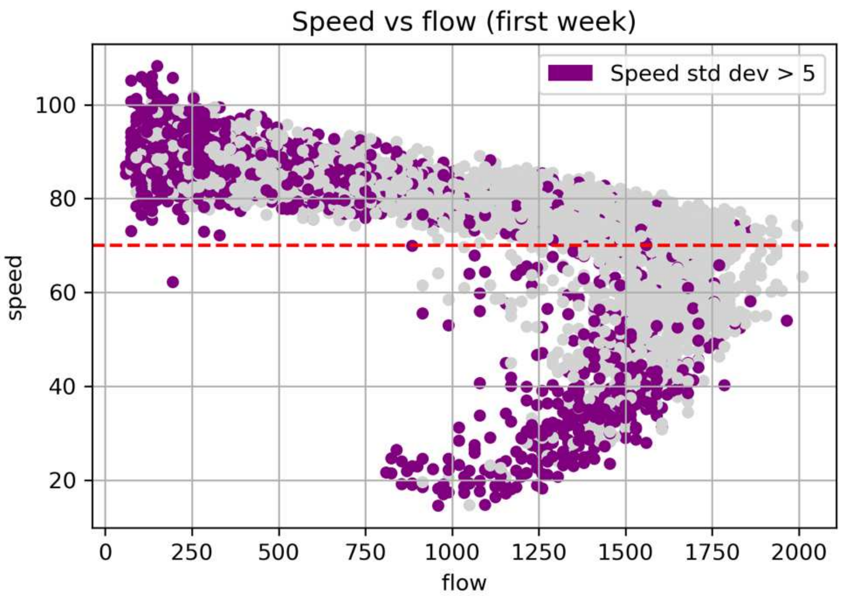

Figure 24 Flow versus speed with standard deviation in purple (speed deviation greater than 5 km / hour)

If we are given a speed value of 80 km/h we know that we are in the uncongested phase. However, the flow can range from 60 to 1600 veh/h. Additional information about the speed standard deviation can help us predict if we are a low or high flow.

Processing INRIX and MCS data (northbound)

A first look at the INRIX data and a comparison with MCS on the timescale of hours. The main purpose here is to look for possible features that can help us predict MCS flow from INRIX speed. There seems to be a two-hour difference in time between INRIX and MCS, minus a 6-minute lag. We added 114 minutes to the INRIX time to be able to compare the two data sets:

```
inrx['timestamputc'] = inrx['timestamputc'] +
datetime.timedelta(minutes=114)
```

Also, in the data, at 2018-10-28 03:00, there is a change to wintertime which has to be dealt with. Here, in the first iteration, look at a subset of that data from 2018-10-01 02:00:00 to 2018-10-21 23:59:59

```
inrx = inrx['2018-10-01 02:00:00':'2018-10-21 23:59:59']
```

Re-sample the data from minutes to hours

a. `inrx60Min['speed'] = inrx.speed.resample('60Min').mean()`
b. `inrx60Min['speedvar'] = inrx.speed.resample('60Min').var()`

Below we see plots of MCS and INRIX hourly average speeds. First over two days and then over a week. INRIX speed is in general lower.

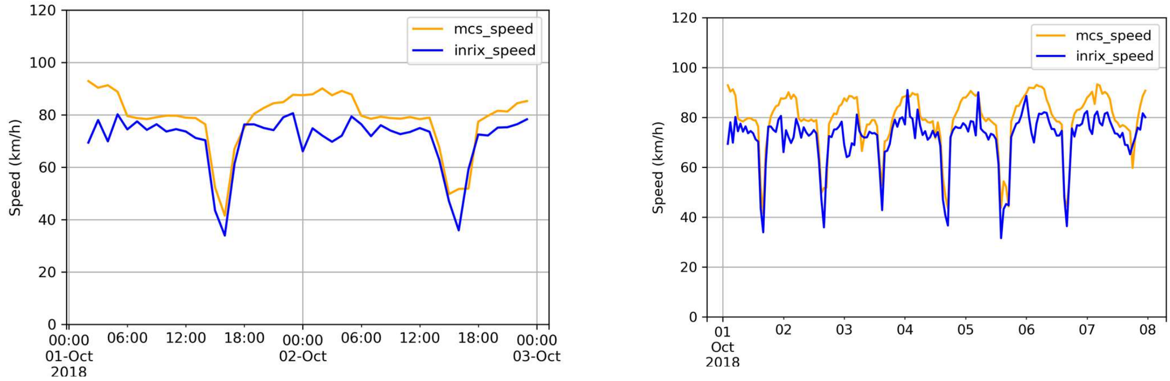

Figure 25 Daily and weekly MCS and INRIX speeds

Plots of speed versus flow using hourly averages are as we have seen previously. We see a clear free-flow phase, with speeds above 70 km/h. The gradient is quite steep in the free-flow phase, compared to the traffic theory in Notley et. al. This makes it easier to predict flow from speed.

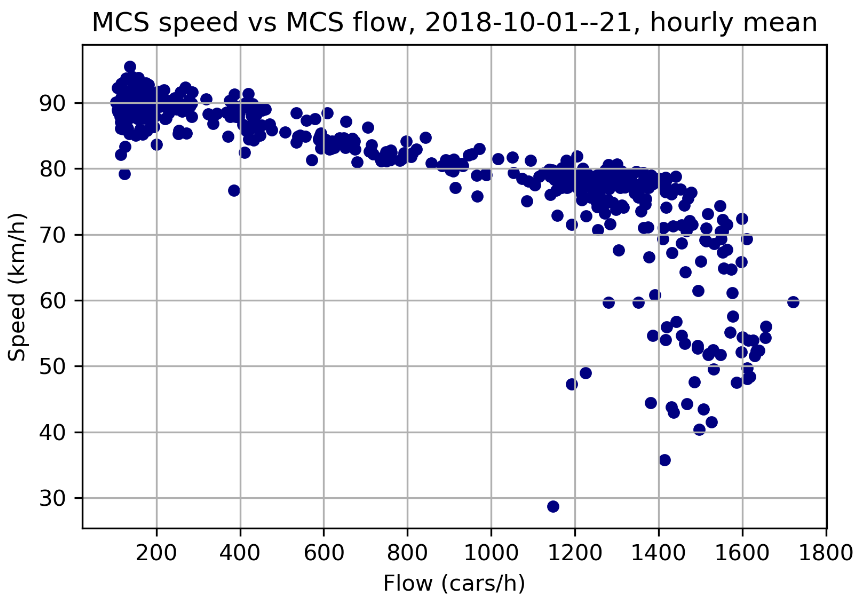

Figure 26 Speed versus flow for MCS data only

If we look at INRIX speed versus MCS flow we see a larger spread of speed values at low flow (top left corner of the plot below).

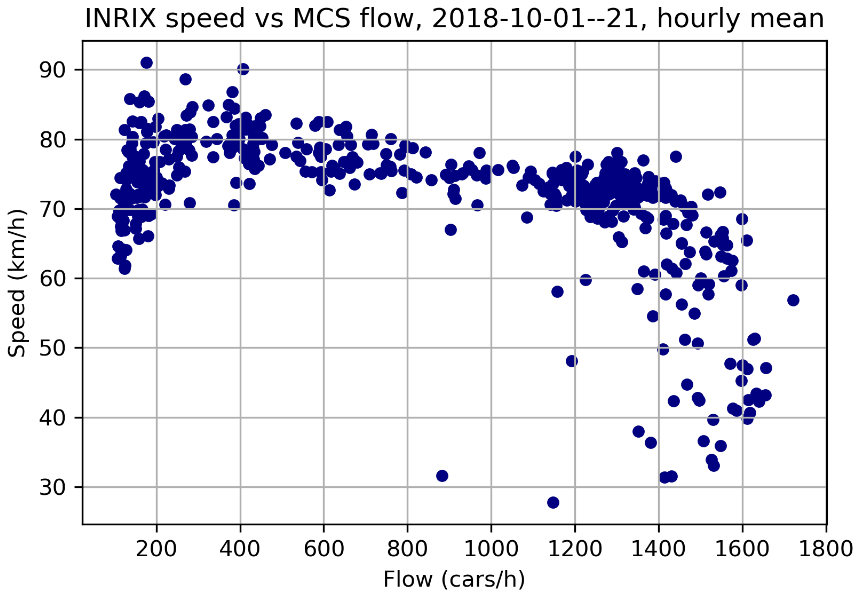

Figure 27 Speed versus flow for MCS and INRIX data

Feature engineering : other INRIX fields

Can speed variance help to predict flow in the uncongested state? The idea presented by Blandin et. al. and described in previous sections is that the speed variance is expected to be high if the flow is low in the uncongested phase. Because drivers are not limited by other traffic and can choose speed freely. At higher flows and still in the uncongested phase the speed variance is expected to be smaller.

In our datasets we don't have the speed of individual cars and we don't have the speed variance. With INRIX we have the average speed of possibly many vehicles per minute. Here the data is aggregated and calculated the mean speed per hour, and at the same time also calculated the variance of the per minute speed values during each hour. Below we see again two scatterplots with INRIX speed versus MCS flow, and this time the colors show speed variance. The first plot shows the full range of speed variance values. The second plot highlights speed variance values above a certain threshold.

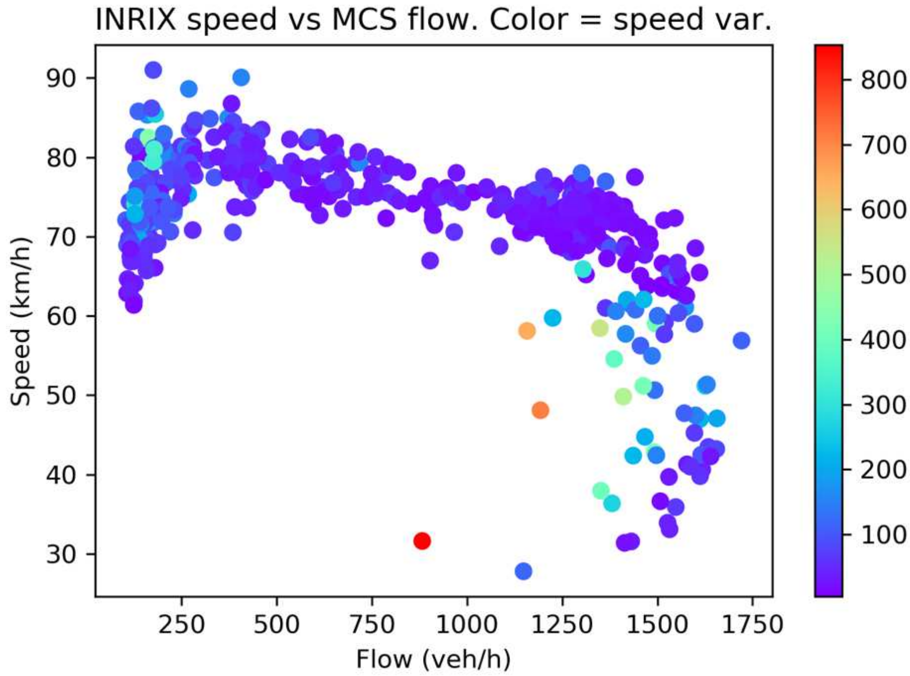

Figure 28 Flow versus speed for MCS and INRIX data (colours = speed variation)

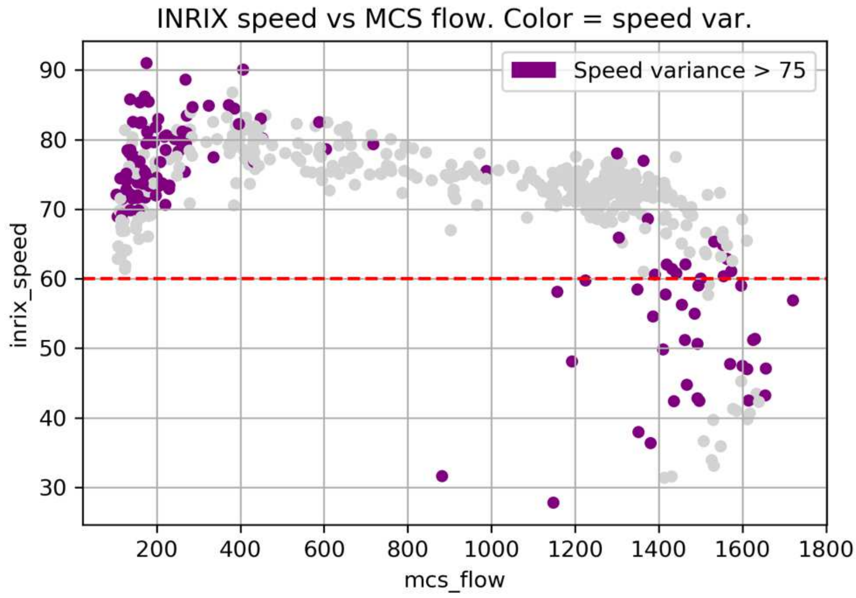

Figure 29 Flow versus speed for MCS and INRIX data (colours indicate the speed variation > 5 only)

We see that the speed variance is higher at low flow in the uncongested phase, top left corner of Figure 29. On this road, the speed ranges from 60km/h to 95km/h during periods with low flow. The variance is larger for the INRIX speeds than for

MCS speeds. And the variance can be high in most of this range. So high variance does not necessarily imply high speed. But the important thing here is the correlation between *flow* and speed variance. The scatterplots below show MCS flow versus INRIX speed variance. First for all traffic, both free-flow and congested and then, more interestingly, with focus on the free-flow state, here identified by a speed greater than 70 km/h.

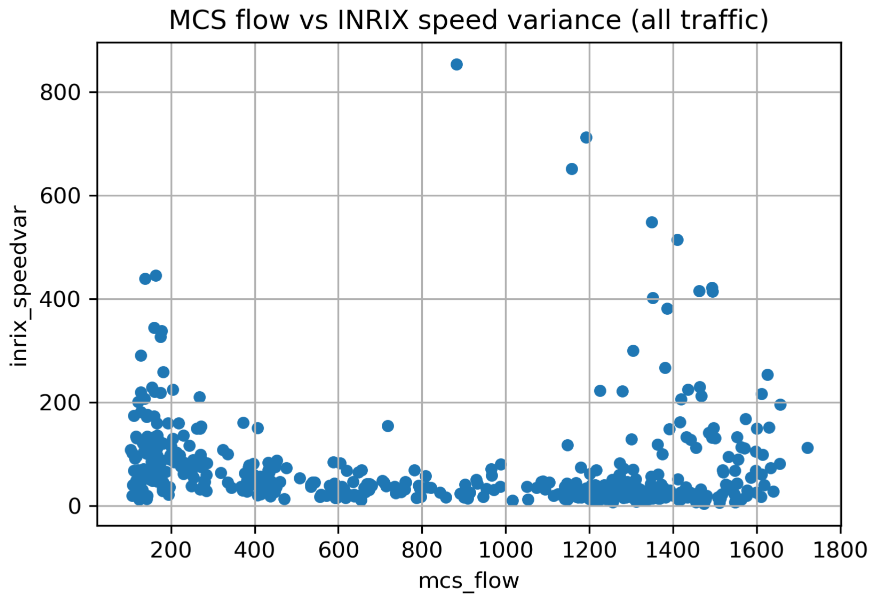

Figure 30 MCS Flow versus INRIX speed (all states)

1. Pearson's correlation: -0.110
2. Spearman correlation: -0.315

A correlation coefficient measures the extent to which two variables tend to change together. The Pearson correlation evaluates the linear relationship between two continuous variables. A relationship is linear when a change in one variable is associated with a proportional change in the other variable. The Spearman correlation evaluates the monotonic relationship between two continuous or ordinal variables. In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate. The Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data, for examples see[2].

[2]https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/supporting-topics/basics/a-comparison-of-the-pearson-and-spearman-correlation-methods/

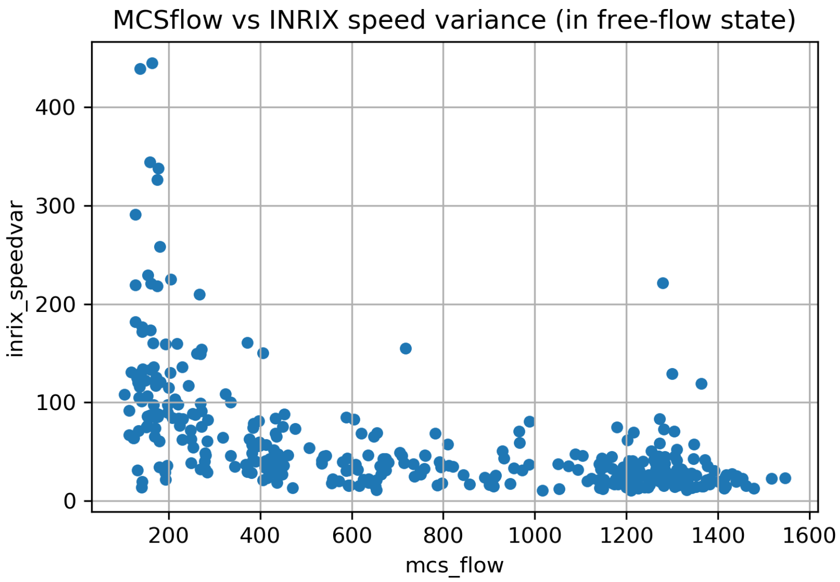

Figure 31 MCS Flow versus INRIX speed (in the free-flow state)

1. Pearson's correlation: -0.533
2. Spearman correlation: -0.644

We see that there is a correlation in the free-flow state. High speed variance indicates low flow (top left). It is still an open issue whether this is sufficient (not necessary) for speed variance to be a useful feature when predicting flows.

Can INRIX reference speed help us to predict flow? *No*, reference: “The free flow speed on the segment for the given day and time”, This field is constant at 60 km/h for all our INRIX data. *Not useful for predicting flow.* Can INRIX speedbucket help us to predict flow? Probably not as speedbucket: Level of congestion, which can have values: 0, 1, 2, 3. The Speedbucket is related to speed rather than flow. The speedbucket value seems to be high when the speed is high enough, irrespective of flow. When re-sampling and aggregating the data from minutes to hours we selected the minimum speedbucket value during each hour.

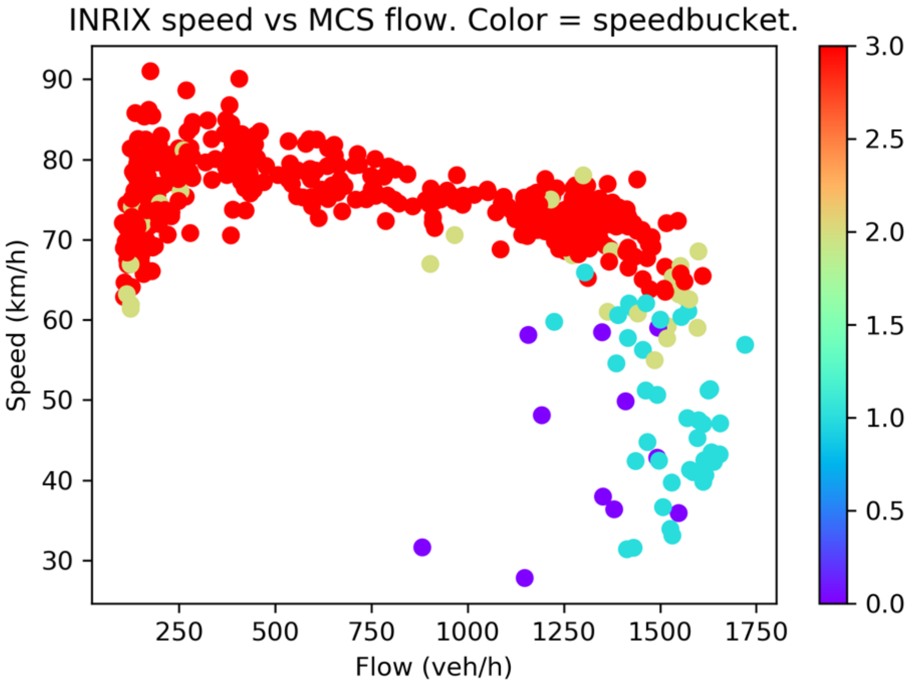

Figure 32: Flow versus speed, colours show the speedbucket field

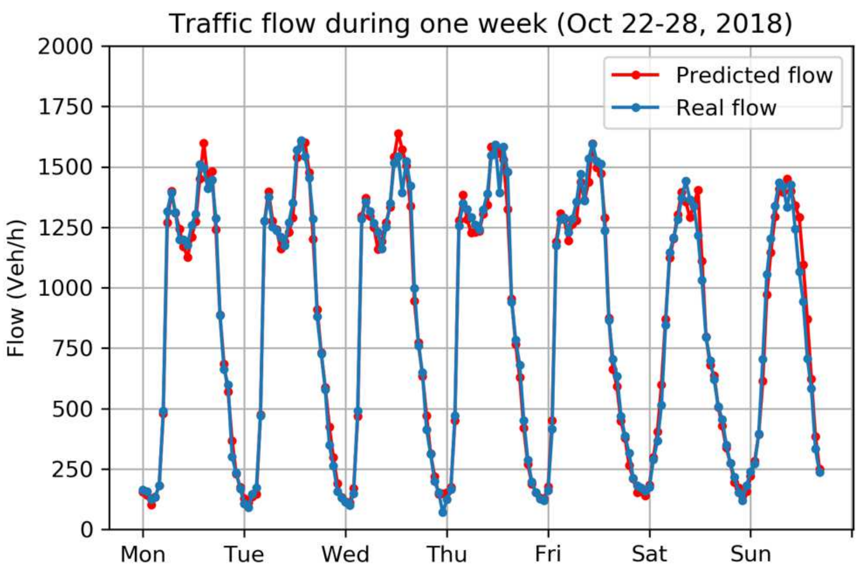

Figure 33: Flow prediction

As can be seen the prediction is very good. Week four is very similar to the previous three weeks on an hourly basis. The mean absolute error is: 35.6, which means that on average the prediction is off by 36 cars. The coefficient of determination, the R^2

score is 0.989. The conclusion from this small and initial study is that traffic flow on a road can be very predictable from week to week. The flow history grouped by hour and weekday can often very well predict the upcoming flows for the next week.

Deep Learning approaches

We also used 5 deep learning methods for estimating the flow for the road data sections we had access to. For a full account of the methods, results and conclusions see [*Hsu2021*]. They are:

1. A univariate neural network
2. A multivariate neural network
3. A neural network with temporal dependency
4. A multivariate neural network with temporal dependency
5. A neural network with spatio-temporal dependency

All the proposed neural networks are based on an architecture called "Wide & Deep Learning," introduced in a 2016 paper [*Cheng2016Wide*]. The architecture enables the neural network to learn both deep patterns through deep neural network layers and simple patterns in data by connecting the inputs directly to the output layer. Each neural network consists of six layers: one input layer, three hidden layers, one concatenate layer and one output layer. The first layer is considering temporaldependencies.an input layer with dimension n equals to the number of input features.

For example, if the model's independent variables are speed and travel time, equals two. The input layer is followed by three densely connected hidden layers containing 120, 60, and 30 neuron nodes, respectively. The number of neuron units in each hidden layer is determined by hyperparameter tuning. The first and third dense hidden layer outputs are fed into a concatenate layer, which merges the outputs of two layers and feeds the concatenation to the output layer. The purpose of the concatenate layer is to provide a short path to the model, through which it can learn simple and undistorted patterns from the layer closer to the input features.

The last layer of the model is an output layer, which densely connects its inputs and produces a predicted value of traffic flow on the road segment as the entire estimator's output. The dimension of the output layer is one for all neural network models. Figure 34 shows the structure of one of the proposed ANN models with 27 input features, including speed, hour, and day features.

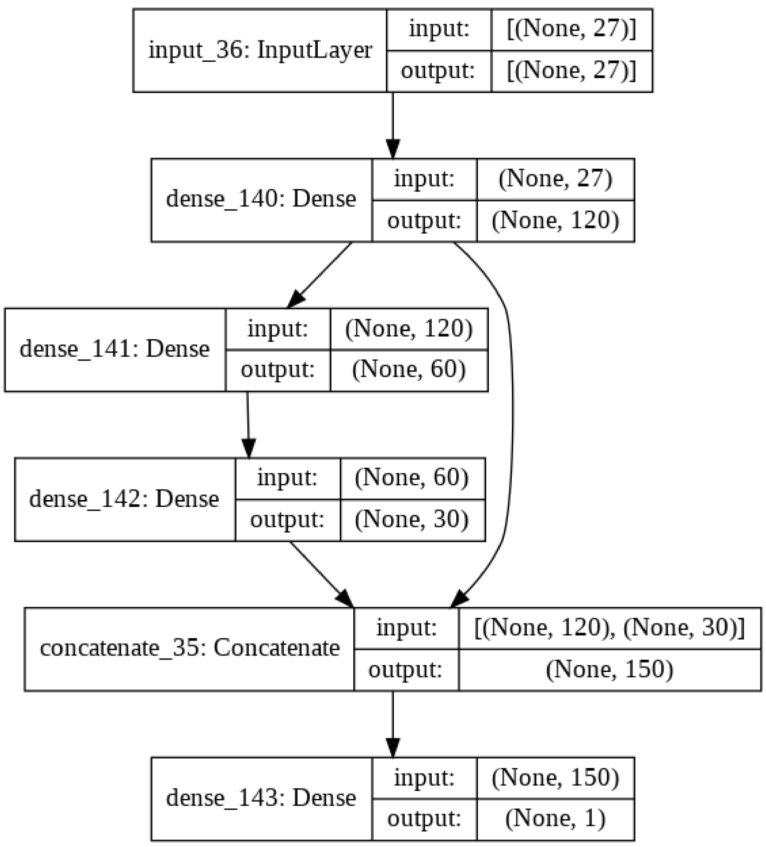

Figure 34: structure of the deep learning model

Two performance measures tend to be used in deep learning, RMSE and MAPE. They are used to evaluate the trained estimation models accuracy when applied to the test datasets. RMSE measures the differences between values predicted by an estimator and the actual value, i.e., the prediction errors, and computes the square root of the average of squared errors. RMSE is also a straightforward measure that has the same unit as the predicted variable, i.e., veh/h in this thesis. On the other hand, MAPE expresses the accuracy as a percentage error by computing the average of absolute ratios of prediction error to the actual value. The MAPE in the project is presented in percentage after multiplying by 100.

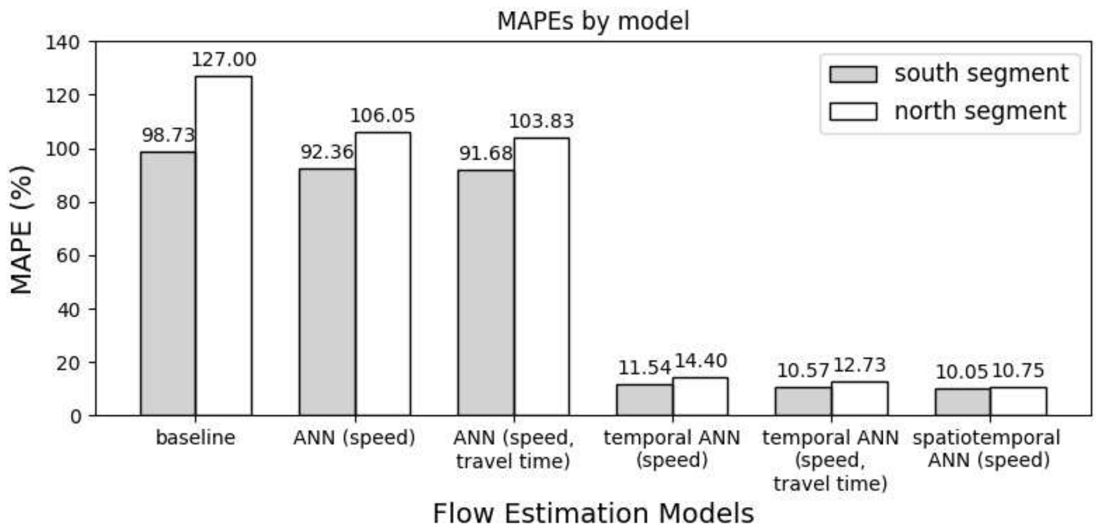

Figure 35: structure of the deep learning model

Figure 35 represents the overview of the estimation performance in MAPE for all models when validated on the south and north test datasets collected in a week, i.e., 22nd to 29th, following the training dataset, i.e., 1st to the 21st. The models in the figure are presented from left to right with increasing complexity. The baseline model on the left is the simplest model with only one input feature, speed, and the spatiotemporal neural network on the rightmost is the most complex, having the most input features and model parameters. As shown in the figure, incorporating additional features, and increasing the model complexity generally can improve the estimation performance on both segments. Again, we refer to the full report [*Hsu2021*] for details on the north and southbound differences as well as the performance metrics.

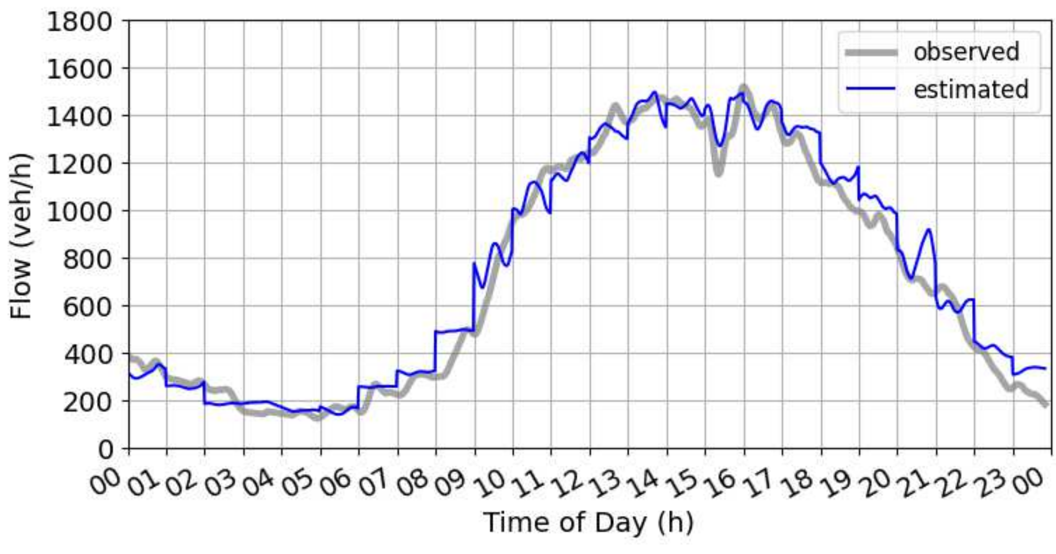

Figure 36: Example prediction from the neural network prediction (5 above)

6 Related work

Nikovski compared non-parametric KNNs, regression trees, locally weighted regression and neural networks to predict the short-term travel times in [*Nikovski2004*]. Duan also used deep neural networks, an Autoencoder to impute/estimate the missing flow data from traffic sensors, and compared its performance with other non-parametric models, e.g., linear regression and back-propagation neural networks [*Duan2016*]. Other non-parametric models include k-nearest neighbours, neural networks, deep learning and so on [*Lint2012*, *LV2015*, *Ma2015*, *Polson2017*, *Zhao2017*, *Li2017*].

Linear regression

Linear regression belongs to the non-parametric methods for solving traffic prediction/estimation problems as both the model structure, the number of independent variables or degree of polynomial, and values of the parameters, the intercept and coefficient of each dependent variable, are determined from the historical data. Linear regression uses historical data to fit a function that characterizes how each covariate, known as the independent variables, influences the outcome variable known as the dependent variable. This linear function is then used to estimate/predict the flow, speed, density, based on unseen data.

Many studies use linear regression as a baseline model to compare with some more sophisticated models, such as neural networks, for traffic flow prediction [*Lint2012*]. Kwon et. al.. use a linear regression model to predict future travel time based on data from loop sensors such as the flow and occupancy, and probe sensors from the departure times and week days [*Kwon2000*]. Nikovski et. al. also used a simple linear regression model with 1 to 3 most recent travel times as input variables to predict the travel time in a short-term future [*Nikovski2004*].

Some studies also adopted linear regression models to estimate traffic states based on partially observed spatial data [*Seo2017*]. For example, Chen et. al. used a linear regression model to estimate the missing data for loop sensors based on traffic data from neighboring sensors around the sensor with missing data [*Chen2007*].

Although linear regression models for traffic prediction are sometimes outperformed by nonlinear models in traffic flow prediction due to the nonlinear nature of traffic flow, its computational efficiency and low memory usage still give it a competitive edge while short execution time is a critical requirement. Moreover, linear regression models sometimes show comparable accuracy in predicting traffic flow as other nonlinear models [*Nikovski2004*]. In this project, linear regression models mainly serve as baseline models for performance comparison with other models. For example, we use linear regression as one of ML models to describe the relation between INRIX data and macroscopic traffic flow in MCS data.

Parametric and non-parametric traffic models can also be combined. Parametric and nonparametric traffic state prediction techniques have previously been developed with different advantages and shortcomings. While nonparametric prediction has shown good results for predicting the traffic state during recurrent traffic conditions, parametric traffic state prediction can be used during nonrecurring traffic conditions, such as incidents and events." In this paper, parametric and nonparametric traffic state prediction techniques are combined through assimilation in an ensemble Kalman filter. For nonparametric prediction, a neural network method is adopted; the parametric prediction is carried out with a cell transmission model with velocity as state." [*Allström2106*]

Traditionally, traffic has been measured with expensive stationary road sensors that provide information about flow, speed and occupancy. These sensors are sometimes referred to as eulerian sensors. Traditional traffic models have been developed based on this data, and therefore often need flow or density as input. But today it is common to collect data from probe vehicles (i.e. GPS-equipped cars and smartphones). Probe vehicle data is sometimes also called mobile data, floating car data (FCD) or Lagrangian data. That data is continuously gathered from the vehicles while driving. The data gives information about speed and traveltime, but it does not typically give any flow information. Given this new type of data, researchers are trying to: (a) adapt or create new traffic models that incorporate probe data; and (b) find methods to derive traffic flow from vehicle probes that measure speed and travel time [*Seo2017*]**.**

One common approach to estimate the traffic flow from probe data is to use a fundamental diagram that provides the relationship between speed and flow for a specific road. Given speed data the flow can, at least in principle, be calculated. But mapping from speed to flow using a fundamental diagram can be challenging, especially in the free-flow phase [*Herrera2010, Blandin2012, Anuar2016, Seo2017*].

Blandin et. al. [*Blandin2012*] studied the empirical relation between point speed and point flow for stationary traffic radars in the San Francisco Bay Area, California. They studied stationary measurements but the goal, in the end, was to assess the feasibility of inferring traffic flow from probe speed. The authors emphasize that in classical traffic flow theory, using a triangular fundamental diagram, velocity is constant at the free-flow speed in the uncongested phase. If the spacing between vehicles is large enough, the drivers will not be constrained by the surrounding vehicles, and so they can travel at free-flow speed. This means that in the uncongested phase, conversion from speed to flow is theoretically impossible as speed is theoretically constant at the free-flow speed. However, Blandin et. al. show with empirical measurements from radar stations that in reality the speed-flow relation can look very different at different roads, flat, increasing linearly, decreasing linearly, non-linear. Blandin et. al. use linear regression to estimate flow from speed, calibrating the model using historical stationary data. They also investigate regression of flow over speed variance. The paper concludes that the proposed methods give reasonably accurate flow estimates, and that the conventional speed-flow method gives significantly more accurate results than the speed variance-flow method.

Using probe data raises a number of additional issues to consider: A basic question with the FD approach is how a well-calibrated fundamental diagram is created. If the objective with probe measurements is to replace expensive stationary measurements, then it doesn't work to use loop-detector data to provide a good FD. A fundamental diagram from another road in the same category might then be used, if it can be assumed to be similar enough. There are also efforts to derive the FD without stationary measurements. For instance, Seo et. al. [*Seo2019*] present a framework for estimating the fundamental diagram without stationary detectors. Instead they use only probe data and information about the jam density.

The aggregation interval of the speed data is often also an important issue. It is also important to consider how representative the speed of the probe vehicles are, compared to all traffic. How large of a share of the vehicles on the road provides speed data (the penetration rate)? Do we mostly get values from a fleet of slow trucks and/or fast driving taxis? Another issue to consider is that probe techniques provide space-mean speed while stationary sensors would give time-mean speed. The characteristics of a specific road segment can also influence the speed and flow estimates (ramps, variable speed limits etc.) compared to measurements at a fixed point.

Herrera and Bayen [*Herrera2010*] propose two methods to incorporate speed measurements from vehicles into flow models for traffic state estimation purposes: a Kalman filtering technique and a Newtonian relaxation method. The latter technique modifies the Lighthill–Whitham–Richards partial differential equation to include a correction term which reduces the discrepancy between the probe measurements and the estimated state. Converting speed measurements into density using a fundamental diagram introduces some errors. However, the paper concludes that despite this error the proposed methods produce accurate estimates.

[*Neumann2013a*] derive traffic volumes from probe vehicle data by applying the speed-flow relationship of the fundamental diagram (using the van Aerde model) on hourly averaged data. Evaluation is done with data from 600 local detectors and a taxi fleet with 4300 vehicles in Berlin, Germany. In [*Neumann2013b*] the authors say that the approach is "more or less applicable, in principle", but the deterministic modelling of the speed-flow relation does not capture the variations in speeds given the same traffic flow. [*Neumann2013b*] proposes a more detailed representation of the fundamental diagram (speed and flow) based on Bayesian networks which also takes into account the dynamic transitions between traffic states over time.

K. A. Anuar estimates traffic volume from probe vehicle data specifically for freeways in the Ph.D. thesis [*Anuar2016*]. Fundamental diagrams is one of three methods explored in the thesis (the other two approaches being: shockwaves and information about the space headway between lead and follower vehicles). The thesis studies four different fundamental diagrams (Greenshields, Underwood, Northwestern, Van Aerde) and data aggregated in 5, 10 and 15 minutes intervals. The results using fundamental diagrams are also presented in the paper [*Anuar2015*]. The probe vehicle data used in this study comes from the Mobile Century project in San Francisco (2008) where GPS data (timestamp and position) was collected from vehicles with recruited drivers. The estimated traffic flows are compared to measured traffic flows from loop detectors. In this case study the Van Aerde-model provides the best results. It gives reasonable estimates of traffic flow rates with good estimates during periods of high flow rates, and with slight underestimation during medium flow rates. Aggregating the data into 15-minutes intervals smoothens the data and gives lower estimation errors. Traffic flow rates are more accurately estimated during congested traffic conditions compared to free-flow conditions in this study.

Road traffic measurements are performed by many different stakeholders: government agencies, municipalities, researchers and commercial companies. The measurements are done for many different reasons, on different time scales, and with different techniques. See for instance Allström et. al. [*Allström2017*] and Sharma et. al. [*Sharma2017*] for an overview of traffic sensors and data collection techniques.

Government agencies and municipalities[3] often measure road traffic for planning of future roads, for operation and maintenance, and for analysis of accident risks and environmental impact. The timescale of interest is often days, months or years. But sometimes measurements are also collected and used in real-time for traffic control and incident detection. The measurement techniques often include inductive loops, radars or pneumatic tubes. The Swedish Transport Administration, “Vägverket”,

[3] "Kartläggning av trafiktekniska mätningar och hur kommuner använder dem", 2017, Marcus Nilsson and Pontus Karlsson, student paper, Lund University. <https://lup.lub.lu.se/student-papers/search/publication/8918020>, viewed 2020-06-16.

measures road traffic on the government owned roads in Sweden[4]. They measure for instance traffic volume, "trafik arbete", expressed in vehicle kilometers, to understand the usage of the road network. In order to monitor traffic trends in the country over time, the Swedish Transport Administration continuously measures traffic on about eighty road sections[5]. These roads are selected to represent the entire state road network in Sweden. Complementary sampling is also done on many other stretches of road.

Several commercial companies, including Google, Waze, Here, TomTom, INRIX and many others, today collect and provide traffic information to drivers (and in some cases also to automotive companies, cities and road authorities). As an example, Google Traffic relies on crowdsourcing from drivers to collect traffic information. Google collects GPS information from phones and calculates the speed of the users on the road. In 2007, Google started to offer live traffic information on top of Google maps. A color code (green, orange, red) is used to highlight the speed of traffic on a road: green means no traffic delays and the darker the red the slower the traffic on the road. Waze also relies on crowdsourcing from drivers to provide a real-time traffic service. They collect map and traffic information, and also allow users to report incidents on the road via a phone app [*WAZE*]. Google bought Waze in 2013. INRIX provides traffic information to road authorities, cities, automotive industries and individuals. They collect traffic data from many sources including road sensors, connected cars and mobile devices [*INRIX*]. Here Technologies is another example of a company that provides map, traffic and location services to both companies and individuals [*HERE*]. Also telecom companies collect data about traffic and mobility and provide services based on the data. In Sweden an example of this is Telia Crowd Insights [*Telia*].

New technology opens up new innovative ways of measuring traffic[6]. This includes video and cameras, Bluetooth and wifi measurements [*Forsman2018*], the use of drones[7], and connected cars talking to each other and the road infrastructure. Techniques that use Bluetooth or wifi are common for Automatic Vehicle Identification (AVI). The Bluetooth or wifi addresses of devices in cars are captured by roadside equipment. The same vehicle can in this way be identified at several locations; and this can be used to calculate travel time.

---

[4] Trafikverket: Vägtrafik- och hastighetsdata <https://www.trafikverket.se/tjanster/trafiktjanster/Vagtrafik--och-hastighetsdata/> , Trafikarbete <https://www.trafikverket.se/tjanster/trafiktjanster/Vagtrafik--och-hastighetsdata/Trafikarbete/>, viewed 2020-06-16.

[5] "Helårsmätpunkter för uppföljning av trafikarbetets förändring på statliga vägnätet" (map of monitoring points in Sweden) <https://www.trafikverket.se/contentassets/5abd840104264f72b7340e481b3771db/tf_punkter.pdf>, viewed 2020-06-14.

[6] BBC, "The technology that could end traffic jams", 2018, <https://www.bbc.com/future/article/20181212-can-artificial-intelligence-end-traffic-jams>, viewed 17 June 2020.

[7] Datafromsky, <https://datafromsky.com/>, viewed 17 June 2020.

Alternative sensor data like vehicle probe data is much cheaper to collect, compared to traditional stationary sensors like radar and inductive loops. Also, with the new sensors a much larger part of the road network can be monitored. But there is an important difference in what type of data the different methods can provide. Vehicle probe data can provide speed and travel time but typically not traffic flow data. If flow data is needed as input to algorithms for traffic control or other calculations, then there is a need to estimate the flow from speed or travel time data. This estimation problem is the focus of the work in this report. Previous research on this estimation problem was surveyed by Seo et. al. [*Seo2017*].

Blandin et. al. [*Blandin2012*] studied the empirical relation between point speed and point flow for 112 stationary traffic radars in the San Francisco Bay Area, California. They studied stationary measurements but the goal, in the end, was to assess the feasibility of inferring traffic flow from probe speed. The authors emphasize that in classical traffic flow theory (using a triangular fundamental diagram) velocity is constant (at the free-flow speed) in the uncongested phase. If the spacing between vehicles is large enough, the drivers will not be constrained by the surrounding vehicles, and so they can travel at free-flow speed. This means that in the uncongested phase, conversion from speed to flow is theoretically impossible as speed is theoretically constant at the free-flow speed. However, Blandin et. al. show with empirical measurements from the radar stations that, in reality, the speed-flow relation can look very different at different roads (flat, increasing linearly, decreasing linearly, non-linear). The figure below comes from the paper by Blandin et. al. [*Blandin2012*]**.**

Kim and Coifman evaluate INRIX speed data by comparing it against concurrent loop detector data **[***Kim2014***]**. They study two months of data from an urban Interstate freeway in Columbus, Ohio, USA. The paper shows that, at a timescale of five minutes, both plots of the INRIX data and corresponding plots derived from loop detector data show similar patterns of congestion. The authors conclude that INRIX data works well for monitoring traffic but they point out three issues with the data: First, INRIX exhibited a latency of about 6 min compared to the loop detector data. Second, INRIX reports speed every minute, but most of the time the reported speed is identical to the previous sample, which indicates that the speed is calculated over a longer time period. Third, the INRIX confidence measures do not appear to reflect the latency or repeated measures. Kim and Coifman note that since the INRIX process is proprietary, there is no way to know if the INRIX data stream includes measurements from the same sensors that the study uses for evaluation. Comparisons between INRIX data and loop detector data should be viewed in this context. In this report we most often assume that the INRIX speed values are based on probe data and calculated from the GPS speed in vehicles.

Sharma et. al. studied INRIX probe data used for traffic operations and safety management in the state of Nebraska, USA [*Sharma2017*]. They evaluated INRIX-data against PVR (per vehicle record) sensordata. There are two main conclusions in their report: First, there is almost always a speed bias between data streaming from probes, the INRIX data and traditional infrastructure-mounted sensors. The average speed bias for real-time data reported in this work was 6.06 mph which is 9.75 km/h. The second conclusion is that the lack of confidence score of 30 real-time INRIX probe data is a critical issue that needs to be considered when doing traffic analysis.

Ahsani et. al. explores the coverage of INRIX real-time data in the state of Iowa, USA, and demonstrates the growth in real-time data over a 4-year timespan [*Ahsani2018*]. A comparison is made with Wavetronix smart sensors to evaluate INRIX's speed data quality. The paper investigates speed bias: the difference in speed values between the INRIX data and the Wavetronic sensors. Some differences are inevitable due to the differences in data collection methods. INRIX and other probe technologies calculate space mean speed; that is the average speed of vehicles over a length of road. Wavetronix, and other stationary road sensors, instead calculate time mean speed; which is the arithmetic mean of vehicles' speed passing a given point. The paper shows that the speed bias may also depend on speed, segment length and time of day. Ahsani et. al. also study how accurate and reliable INRIX is when it comes to detecting congestion (both recurring and non-recurring).

In addition to factors such as speed and density, which affect flow rate based on traffic theory, studies also found that the relations between flow and speed/density are in fact not static but changes with time, ie, dynamic. Dervisoglu et. al. found that the maximum flow rate of a road segment not only changes on different days, but also changes before and after the congestion occurs under the same critical density, which means the flow-density relation in a road segment [*Dervisoglu2009*]. Duan also found that the flow rate patterns in weekends are different from patterns in weekdays, and using the data with different temporal factors when training could improve the performance of flow data imputation [*Duan2016*]. In this work, we also observed that the relation between flow and speed is different between daytime and night, i.e., a flow drop is observed under the same speed between daytime and night. Therefore, including the temporal factor of daytime/night into our feature vector should help the regression models to estimate the flow rate more precisely.

Previous studies showed that traffic state parameters such as speed and flow rate of road links have spatial correlation with each other, which means that there are dependencies between different road links in the same road network and their traffic states are influenced by their neighbors [*Ermagu2017*]**.** Based on these observations, various studies have utilized the spatial dependency between road links in the problems of traffic state estimation and prediction. For example, Chen estimated/imputed the missing flow rate data of the detector by using data collected from neighbor detectors (in the same timestamp) based on the linear relations of flow rate between the center and neighbor detectors [*Chen2003*]. Among spatial relations between road links, strong positive correlations were observed between the links and their upstream/downstream links [*Ermagu2017*]**.** For example, Duan et. al. found that the accuracy of their flow estimation model could be improved by including the data from upstream and downstream road segments instead of only using the data from a single road segment as the input feature [*Duan2016*]. Moreover, the spatial correlation between road links, especially the positive correlations between target links and their immediate upstream/downstream links, are widely used to solve the traffic prediction and estimation problems [*Zhang2019, Ermagu2017]*. In this work, we also observed a high correlation of INRIX data between south and north road segments at the same time step (correlation > 0.95). Therefore consider the spatial factor by including INRIX data from the adjacent road segment into our model. By doing so we expect to improve the accuracy of estimation since the data from the upstream/downstream segments may provide extra information about traffic states based on the spatial correlation.

A 30-page white paper published by the UK TRL research lab exemplifies the most important traffic flow relationships [*Notley2009*]. Non-linear density-flow, flow-speed, density-speed relationships are explained. In the flow-speed cases traffic *phases* are given, free-flow and breakdown. Some plots show the maximum or optimal traffic situations using the data to indicate capacity for example. Shockwaves, or the back propagation of slowing down, are exemplified on the busy M25 ring road around London. Comparison of time of day effects are shown as histograms (1 per day). This report is directly relevant to our work, in that we will recreate the plots for at least the MCS data. It differs in that it only uses loop sensors, whereas we utilize both loop and probe data. It differs from us in that we use floating car data as well and apply ML methods too. An introductory *textbook* on traffic flow is [*Elefteriadou2014*].

There are also methods that estimate flow from mobile measurements without using a fundamental diagram. These methods often either (a) combine mobile measurements with vehicle counts and other limited stationary data [*Seo2017, Coifman2003, Astarita2006, Qiu2010, Bekiaris-Liberis2016, anthawichit2003, Sekula2017*] or (b) make use of more advanced mobile data (than just speed and traveltime), for instance spacing information [*Seo2015a, Seo2015b, Anuar2016*].

7 Potential future directions

Spatial-temporal correlations

Mining spatial and temporal information in the spatial-temporal correlation datasets is a prevalent topic in traffic data imputation and prediction. We did not in detail the temporal and spatial correlation between neighboring data points in the time-space domain for traffic flow estimation and imputation. We did capture the traffic condition's dependency on the hour/weekday and location in the historical data and used this information to infer the traffic flow based on the real-time INRIX measures. However, more information is hidden in an unobserved/missing data point's neighboring data in the time-space domain. For example, a road segment's traffic flow could be heavily affected by the traffic conditions from its upstream and downstream road segments. Moreover, traffic flow in a time slice could have strong correlations with traffic conditions in its previous and later few time slices. Therefore, full use of spatial-temporal information could help us improve the accuracy of traffic flow estimation, imputation, or prediction. One possible direction for future research is to collect a comprehensive dataset containing INRIX and MCS data collected from many road segments on the road or in a road network during an extended period.

One could use sophisticated deep learning methods such as Recurrent Neural Network (RNN), CNN, or Graph Convolutional Network (GCN) to extract the temporal and spatial correlations between traffic flow and INRIX's measures from the dataset containing rich spatial-temporal information. The extracted spatial-temporal correlations and real-time INRIX measures could be solely used for estimating the traffic flow on a road segment, like what we did in this work, or could be used together with other fixed sensordata for estimating, imputing, or predicting the traffic flow.

Traffic Flow Prediction

Predicting short-term traffic flow is a challenging but more valuable task for traffic control ITS and transportation planning. Therefore, one possible future direction is extending the approach proposed in this report to work for traffic flow forecasting. Most studies use features solely from the fixed-location sensor's measures to predict traffic flow to the best of the author's knowledge. At the same time, very few of them took alternative data sources, e.g., mobile data, into account when it comes to traffic prediction. The quickest way to extend our approach for traffic forecasting is using flows in the short-term future, e.g., 15 minutes, as labels to train the neural network models, and then using the models to predict the traffic flow based on the current INRIX measures. However, a typical prediction approach will be extracting the temporal correlations from the historical data using some time-series models, e.g., RNN or Long Short-Term Memory (LSTM), and then using the extracted correlation for prediction based on the recent INRIX data from the current and previous time slices. Some recent studies also captured the spatial correlation and the temporal correlation using GCN to make full use of both spatial and temporal information in a road network for traffic forecasting. No matter which method one uses to extract the temporal and/or spatial correlations between INRIX measures and the traffic flow, INRIX data as an additional informationsource should help improve the prediction accuracy when used together with the typical fixed sensor's measures. Besides, we could also use INRIX alone for predicting the short-term traffic flow when other data sources are not available, e.g., the fixed detector malfunctions.

Using density as a traffic control parameter

By density we mean the number of vehicles per road length. Some reasons for preferring traffic density over flow or average speed are:

1. Density is intuitive, we can think of the number of vehicles per road section
2. Velocity is dependent on the vehicle type:
    - Trucks are slower than cars
    - Density is independent of the vehicle type.
3. Density can be compared between road sections, countries, etc.
4. Average velocity and flow can be difficult or sensitive to local conditions
5. Density is a better input to emissions as the #vehicles / area can be correlated to the CO_2 or heavy NOX particles. See the other TENS reports.

In some cases, extra inference is needed to obtain queue buildup between sensors, as queues can accumulate very quickly, i.e. at higher resolutions than the cross-sectional traffic. Space does not permit its derivation here, but density changes are obtained from the PDE relationship between flow-density-velocity.

Alternative prediction methods

Given the periodicity and relatively regular pattern of flow and speed using spectral techniques is a possibility. The better known one is Fourier analysis and its associated Fourier transform (FT). It is a mathematical transform that decomposes functions depending on space or time into functions.

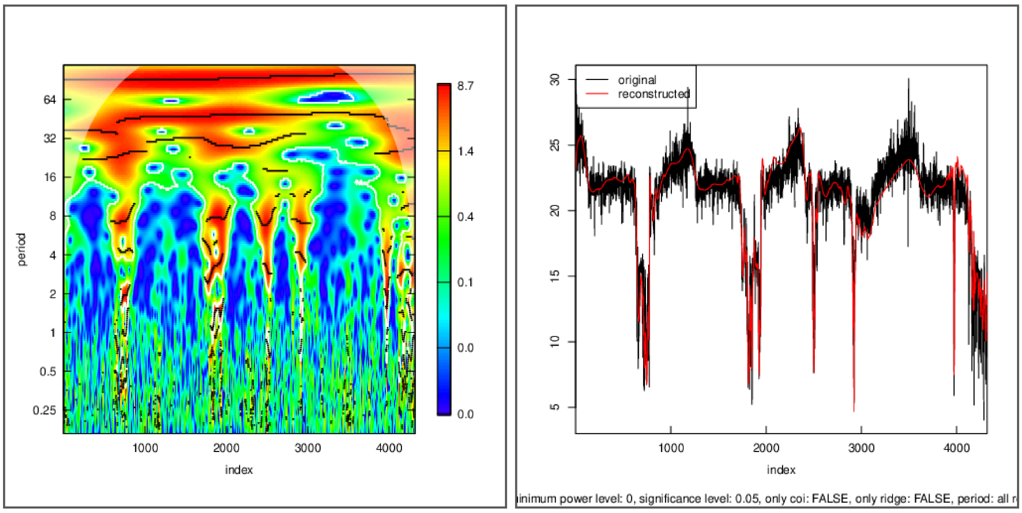

Figure 15: Wavelet decomposition of speed (left), subsequent reconstruction (right)

Wavelets are termed a "brief oscillation". It is a wave-like oscillation with an amplitude that begins at zero, increases or decreases, and then returns to zero one or more times. A taxonomy of wavelets has been collated which is based on the number and direction of the pulses. Wavelets have properties that make them useful for signal processing, time series, coding media and so on.

8 Conclusions

Traffic flow is a complex phenomenon. Although relatively predictable on longer time scales as we have plotted, as the time scales shorten the prediction becomes more difficult. From a driving point of view, individual drivers can make irrational maneuvers that affect the entire flow of a lane, road or city.

Using regression and the fundamental diagrams time-of-day is a good predictor for flow. The average flow per hour during the first three weeks of the data, predicts well the flow per hour during the fourth week. Speed variance might be a useful feature for improving the prediction of flow from speed during the free-flow phase. A first test with a simple linear regression with two line-segments, separating the free-flow and congested state, gives a good score (0.89 R^2). Also using neural networks show similar results, albeit at higher complexity, more data, but better opportunities for using feature engineering.

In this report (with background in the first report) we have looked at the predictability of flow in MCS and INRIX systems. Then we take one step further and try to infer the flow in the complex MCS system from the potentially easier to gather floating or Bluetooth sensors. Pros and cons are available in the other TENS reports.

9 References

[*Bando1995*] Bando M, Hasebe K, Nakayama A, Shibata A, Sugiyama Y. *Dynamical model of traffic congestion and numerical simulation*. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics. 1995 Feb;51(2) 1035-1042. doi:10.1103/physreve.51.1035. PMID: 9962746.

[*Bertsimas2018Travel*] Dimitris Bertsimas, Arthur Delarue, Patrick Jaillet and Sebastien Martin, *Travel Time Estimation in the Age of Big Data*, Operations Research Center, Massachusetts Institute of Technology May 2018, http://web.mit.edu/jaillet/www/general/travel-time-18.pdf

[*Bhouri2019*] Neila Bhouri, Maurice Aron, Habib Haj Salem. *A Data-driven Approach for Estimating the Fundamental Diagram*. Promet - Traffic & Transportation, 2019, 31 (2), pp. 117-128. ff10.7307/ptt. v31i2.2849ff. hal-02110696v2

[*Bertsimas2018*] Dimitris Bertsimas, Arthur Delarue, Patrick Jaillet and Sebastien Martin, *Travel Time Estimation in the Age of Big Data*, Operations Research Center, Massachusetts Institute of Technology May 2018, http://web.mit.edu/jaillet/www/general/travel-time-18.pdf

[*Blandin2012*] Blandin, S., Salam, A., Bayen, A.M., 2012. *Individual speed variance in traffic flow: analysis of Bay Area radar measurements*, in: Transportation Research Board 91st Annual Meeting.

[*Cheng2016Wide*] T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispiret al., *Wide and deep learning for recommender systems*, in Proceedings of the 1st workshop on deep learning for recommender systems, 2016, pp. 7–10

[*Cosar2018*] Cosar Ghandeharioon, *An evaluation of deep neural network approaches for traffic speed prediction*, KTH EECS, 2018, Thesis.

[*Duan2016Efficient*] Y. Duan, Y. Lv, Y.-L. Liu, and F.-Y. Wang, *An efficient realization of deep learning for traffic data imputation*, Transportation research part C: emerging technologies, vol. 72, pp. 168–181, 2016.

[*Elefteriadou2014*] An introduction to Traffic Flow Theory. Springer, 2014.

[*Forsman2015*] Åsa Forsman, Anna Vadeby, David Gundlegård and Rasmus Ringdahl, Utvärdering *av hastighetsmätningar med blåtandssensorer Jämförelse med data från MCS (Motorway Control System),* 2015, VTI rapport 969, Link.

[*Helbing2017Microscopic*] Dirk Helbing, *From microscopic to macroscopic traffic models*. 10.1007/BFb0104959, 2007. https://arxiv.org/abs/cond-mat/9806171.

[*Gundlegård2016*] David Gundlegård, Clas Rydergren, Nils Breyer and Botond Anti, *Travel demand estimation and network assignment based on cellular network data*. Computer Communications, 95:29–42, 2016.

[*Gundlegård2020*] David Gundlegård. *Förstudie kring användning av probedata för skattning av tidsvarierande reseefterfrågan och trafiktillstånd.* TENSE project, 2020-11-30.

[*Ghandeharioon2018Evaluation*] Cosar Ghandeharioon, *An evaluation of deep neural network approaches for traffic speed prediction*, KTH EECS, 2018, Thesis.

[*Gundlegård2016Travel*] David Gundlegård, Clas Rydergren, Nils Breyer and Botond Rajna. *Travel demand estimation and network assignment based on cellular network data*. Computer Communications, 95:29–42, 2016.

[*Gühnemann A. et. al*] Astrid Gühnemann, Ralf-Peter Schäfer, Schäfer Kai-Uwe Thiessenhusen and Peter Wagner. *Monitoring traffic and emissions by floating car data*, 2004.

[*Gühnemann2004*] Astrid Gühnemann, Ralf-Peter Schäfer, Schäfer Kai-Uwe Thiessenhusen and Peter Wagner. *Monitoring traffic and emissions by floating car data*, 2004.

[*ITF2015*] International Transport Forum, *The ITF transport outlook 2021*[8], link (previous years also available).

[*Kang2019*] Kang, *Traffic Density Estimation Based on Vehicle Speed Profile Data,* Extended Abstract at the 98th Traffic Board Meeting, 2019.

[*Kim2014*] Seoungbum Kim, Benjamin, Coifman, *Comparing INRIX speed data against concurrent loop detector stations over several months*, Transportation Research Part C: Emerging Technologies. Volume 49, 2014, Pages 59-72, ISSN 0968-090X, https://doi.org/10.1016/j.trc.2014.10.002.

[8] Previous years are also available.

[*Jing2016Development*] Boyu Jing, Lin Wu, Hongjun Mao, Sunning Gong, Jianjun He, Chao Zou, Guohua Song, Xiaoyu Li, and Zhong Wu. *Development of a vehicle emission inventory with high temporal–spatial resolution based on NRT traffic data and its impact on air pollution in Beijing part 1: Development and evaluation of vehicle emission inventory*. Atmospheric Chemistry and Physics, 16(5):3161–3170, 2016.

[*Jing2016*] Boyu Jing, Lin Wu, Hongjun Mao, Sunning Gong, Jianjun He, Chao Zou, Guohua Song, Xiaoyu Li, and Zhong Wu. *Development of a vehicle emission inventory with high temporal–spatial resolution based on NRT traffic data and its impact on air pollution in Beijing part 1: Development and evaluation of vehicle emission inventory*. Atmospheric Chemistry and Physics, 16(5):3161–3170, 2016.

[*Hsu2021*] Pei-Lun Hsu, *Data-driven macroscopic traffic prediction from microscopic measurements*, KTH, 2021, report.

[*Kang2019*] *Traffic Density Estimation Based on Vehicle Speed Profile Data,* Extended Abstract at the 98th Traffic Board Meeting, 2019.

[*Kim2014Comparing*] Seoungbum Kim, Benjamin, Coifman, *Comparing INRIX speed data against concurrent loop detector stations over several months*, Transportation Research Part C: Emerging Technologies. Volume 49, 2014, Pages 59-72, ISSN 0968-090X, https://doi.org/10.1016/j.trc.2014.10.002.

[*Hooper2014*] Hooper, E., Chapman, L. & Quinn, A. *The impact of precipitation on speed–flow relationships along a UK motorway corridor.* Theoretical Applications of Climatology 117, 303–316 (2014) https://doi.org/10.1007/s00704-013-0999-5.

[*Marsh2012*] I. Marsh, *VANET communication: a traffic flow approach*, PIMRC 2012.

[*Marsh2021*] I. Marsh, *End of traffic queue detection*, paper in progress.

[*MarshTENS2021I*] Ian Marsh and Pei-Lun Tsu. D*ata-driven traffic flow: A state-of-the-art report*, RISE report 2021:93, ISBN 978-91-89385-83-2 (this report).

[*MarshTENS2021II*] Ian Marsh, Henrik Abrahmsson, Pei-Lun Tsu. *Data-driven traffic flow, predicting macro flows from micro measurements*.

[*Nagel2003*] Kai Nagel, Peter Wagner, Richard Woesler. Still Flowing: *Approaches to Traffic Flow and Traffic Jam Modeling*. Operations Research 51(5):681-710. https://doi.org/10.1287/opre.51.5.681.16755

[*Notley2009*] Speed, Flow and Density of Motorway Traffic, TRL (Transport Research Laboratory), 2009.

[*Notley2009*] S. O. Notley, N. Bourne, N.B. Taylor. *Speed, Flow and density of motorway traffic*. UK TRL Insight report INS003, pdf (readable report on the fundamentals of traffic flow with examples).

[*Polson2016Deep*] Nicholas Polson, Vadim Sokolov. *Deep Learning Predictors for Traffic Flows,* paper.

[*Polson2015*] Polson, Nicholas; Sokolov, Vadim. *Bayesian analysis of traffic flow on interstate I-55: The LWR model*. Ann. Appl. Stat. 9 (2015), no. 4, 1864–1888. doi:10.1214/15-AOAS853. https://projecteuclid.org/euclid.aoas/1453993096.

[*Raadsen2016*] Mark P.H. Raadsen, Michiel C.J., Bliemer Michael and G.H. Bell. *An efficient and exact event-based algorithm for solving simplified first order dynamic network loading problems in continuous time*. Transportation Research Part B, 92:191–210, 2016.

[*Reginbald2018*] Reginbald Ivarsson, Jón Scalable System-Wide Traffic Flow Prediction Using Graph Partitioning and Recurrent Neural Networks, KTH EECS, 2018, Thesis.

[*Reim2012Analysis*] Erich Reim, *Analysis and visualization of historical traffic data collected on the Stockholm highway system*, 2013-10-30.

[*Reginbald2018*] Reginbald Ivarsson, Jón, Scalable *System-Wide Traffic Flow Prediction Using Graph Partitioning and Recurrent Neural Networks*, KTH EECS, 2018, Thesis.

[*Seo2017*] Traffic state estimation on highway: A comprehensive survey, Annual Reviews in Control. doi: 10.1016/j.arcontrol.2017.03.005.

[*Seo2019Fundamental*] Toru Seo, Yutaka Kawasaki, Takahiko Kusakabe, Yasuo Asakura, *Fundamental diagram estimation by using trajectories of probe vehicles*, Transportation Research Part B: Methodological, Volume 122, 2019, Pages 40-56, ISSN 0191-2615, https://doi.org/10.1016/j.trb.2019.02.005, http://www.sciencedirect.com/science/article/pii/S0191261518303527.

[*Sigurdsson2018Road*] Thorsteinn Thorri Sigurdsson, *Road traffic congestion detection and tracking with Spark Streaming Analytics*, KTH EECS, 2018, Presentation, Thesis.

[*Sharma2017Evaluation*] Sharma, V. Ahsani, and S. Rawat, “Evaluation of opportunities and challenges of using inrix data for real-time performance monitoring and historical trend assessment,”Reports and White Papers, vol. 24, 2017.

[*TrafficFlow*] Wikipedia page.

[*Trafa2022*] Transport Analysis is a government agency charged with providing decision-makers in the sphere of transport policy with sound and relevant policy advice.

[*Tsanakas2019*] Tsanakas N, Ekström J, Gundlegård D, Olstam J, Rydergren C, *Data-driven network assignment*, Swedish National Transport Conference, Linköping, Oct 2019.

[*Tsanakas2019-Lic*] Tsanakas, Nikolaos, Emission estimation based on traffic models and measurements, 2019, Licentiate thesis, monograph. https://doi.org/10.3384/lic.diva-155771

[*UN2021*] United nations, *The 17 goals, sustainable development*, accessed:2021-05-16. Available: https://sdgs.un.org/goals.

[*Williams2003Modelling*] B. M. Williams and L. A. Hoel, *Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results*, Journal of transportation engineering, vol. 129, no. 6, pp. 664–672, 2003.

[*Yatskiv2013*] I. Yatskiv, A. Grakovski, El Yurshevich, *An Overview of Different Methods Available to Observe Traffic Flows Using New Technologies*, NTTS 2013.

[*Yperman2007*] I. Yperman. *The link transmission model for dynamic network loading*. Ph.D. Thesis, Katholieke Universiteit Leuven, 2007.

[*Yang2019Data-driven*] Shu Yang Yao and Jan Wu, *Data-Driven Approaches for Estimating Travel Time Reliability*, 10.1016/B978-0-12-817026-7.00004-7, 2019.

[*Yang2019*] Shu Yang Yao and Jan Wu, *Data-Driven Approaches for Estimating Travel Time Reliability*, 10.1016/B978-0-12-817026-7.00004-7, 2019.

[*Zhan2016*] Xianyuan Zhan, Yu Zheng, Xiuwen Yi and Satish V. Ukkusuri. *Citywide traffic volume estimation using trajectory data*. IEEE Transactions on Knowledge and Data Engineering, 29(2):272–285, 2016.

[*Zainab18Short-term*] Zainab Abbas, Ahmad Al-Shishtawy, Sarunas Girdzijauskas, Vladimir Vlassov, *Short-Term Traffic Prediction Using Long Short-Term Memory Neural Networks,* IEEE International Congress on Big Data 2018, paper.