Project

AI in Transportation

Author: Johanna Schaefer

17.10.2025

1. Introduction

In modern transport systems, continuous collection of traffic data is essential for providing real-time information on speed and traffic flow. Sensors on road portals provide important data, but if individual sensors fail or provide incomplete data, this can affect the accuracy of predictions. The central question of this work is therefore: How well can missing sensors be compensated for by data from other sensors in the same portal or from neighbouring portals? To answer this question, a regression approach is used in which the measured values of a target sensor are predicted on the basis of neighbouring sensors. The study examines whether sensors within the same portal provide better prediction accuracy than sensors from a neighbouring portal.

2. Descriptive analysis

To begin the project of creating predictive models of the data provided, the data needs to be analysed.

2.1. Data description

The data set used for this project consists of speed and traffic flow data from several sensors on the motorway near Stockholm. The data comes from 29 sensors belonging to 8 different portals on a section heading south. Every sensor measures the speed and flow in one line of the motorway. There are inflows and outflows within this section between the portals as can be seen in figure 2.1.

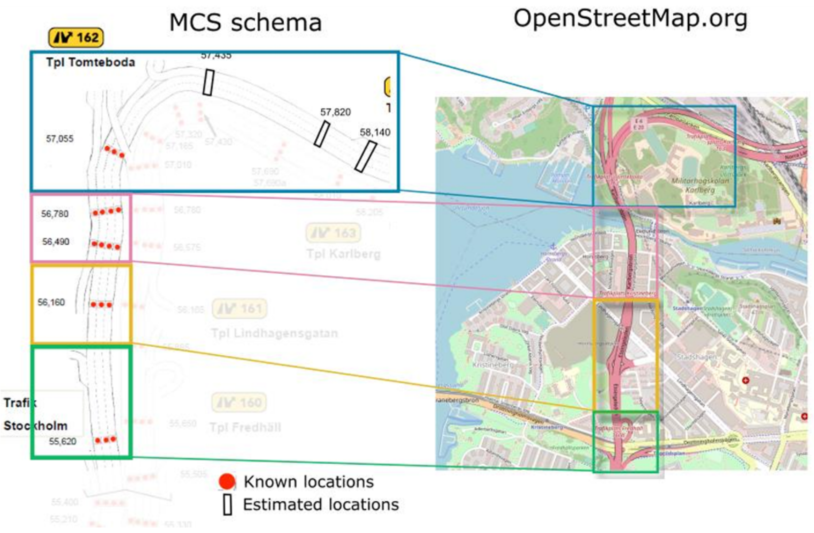


Figure 2.1: Overview motorway section and portals

Table 2.1 shows the portals and the correspondings sensors.

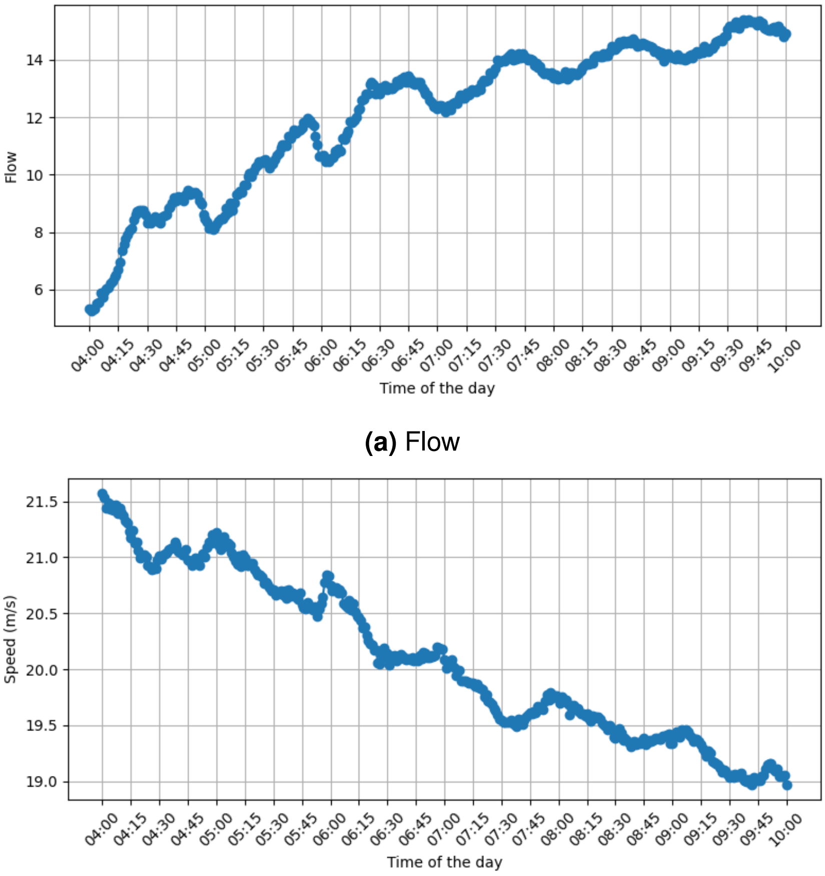
Table 2.1: Portals and sensors

Portal	sensors
E4S 55620	751, 1254, 1076
E4S 56160	539, 536, 740
E4S 56490	351, 4873, 153, 902
E4S 56780	543, 534, 4872, 1079
E4S 57055	353, 1443, 749
E4S 57435	4430, 4429, 4437
E4S 57820	4427, 4436, 4428, 4439
E4S 58140	4494, 4473, 4495, 4472, 4496

Over a total of 214 days, speed and flow data were recorder between 4 AM and 10 AM with a temporal resolution of one minute. Speed is measured in $\mathrm{m/s}$, while flow is quantified as the number of vehicles per minute.

2.2. Speed and Flow over the morning peak

Between 9 AM and 10 AM, the flow of vehicles shows a generally increasing trend, indicating a continued buildup of traffic volume as the morning progresses. The rate at which it in increasing is slowing down when approaching the 10 AM. In contrast, the average speed tends to decline over the same period, reflecting growing congestion. This inverse relationship between flow and speed is characteristic of peak-hour dynamics, where higher vehicle density leads to reduced travel speeds.



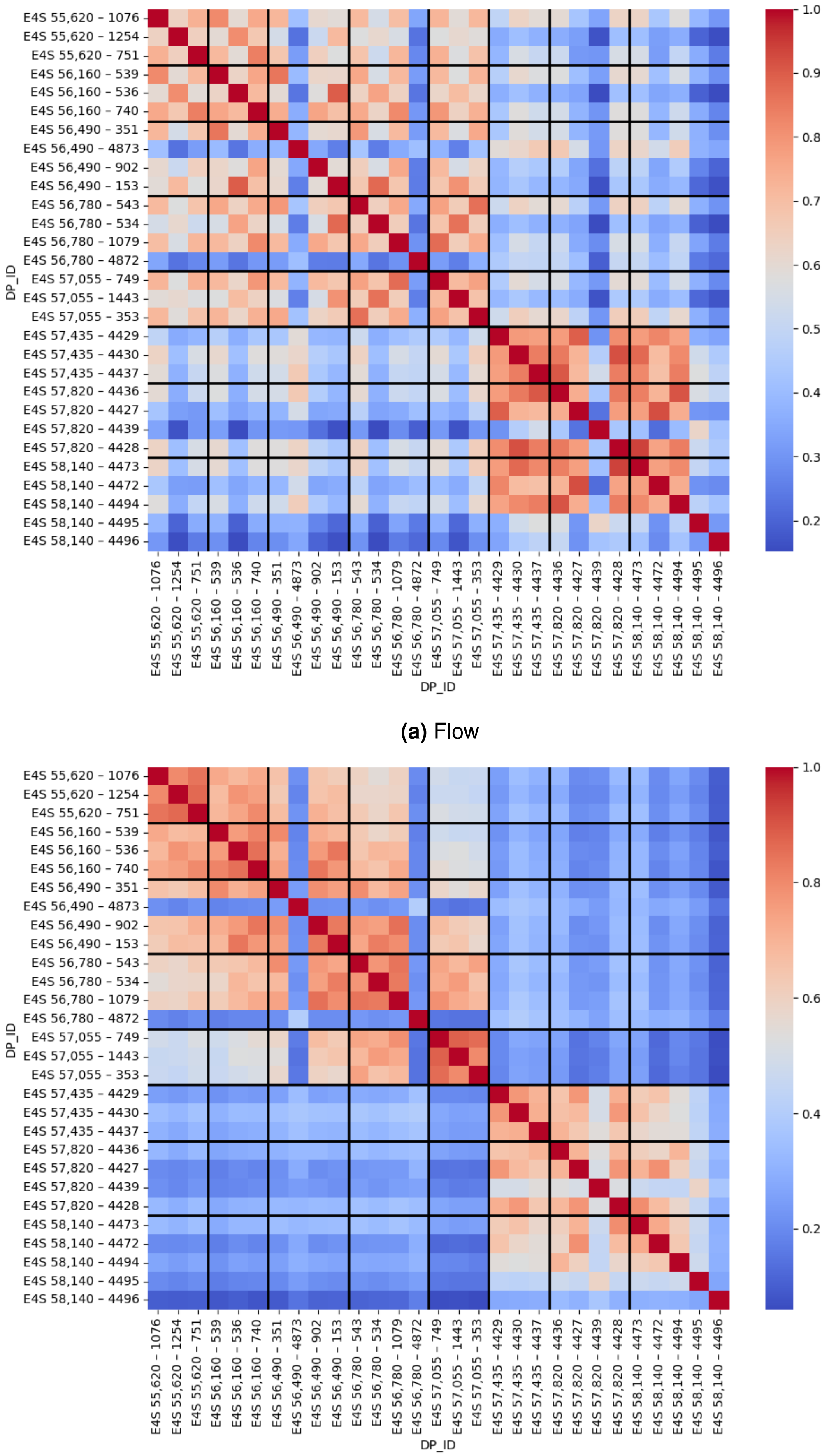
(a) Flow

(b) Speed

Figure 2.2: Flow and Speed over the day

2.3. Correlation

Furthermore, the correlation between different sensors was analysed for both speed and flow. It is important to note that no time lag was considered in this initial assessment. In addition to variations in speed and flow across sensors, spatial separation plays a significant role: the greater the distance between sensors, the more pronounced the temporal offset and divergence in observed traffic patterns. In figure 2.3 the sensors are displayed grouped by the portals they are part of. The thicker black lines indicate the portals.



(a) Flow

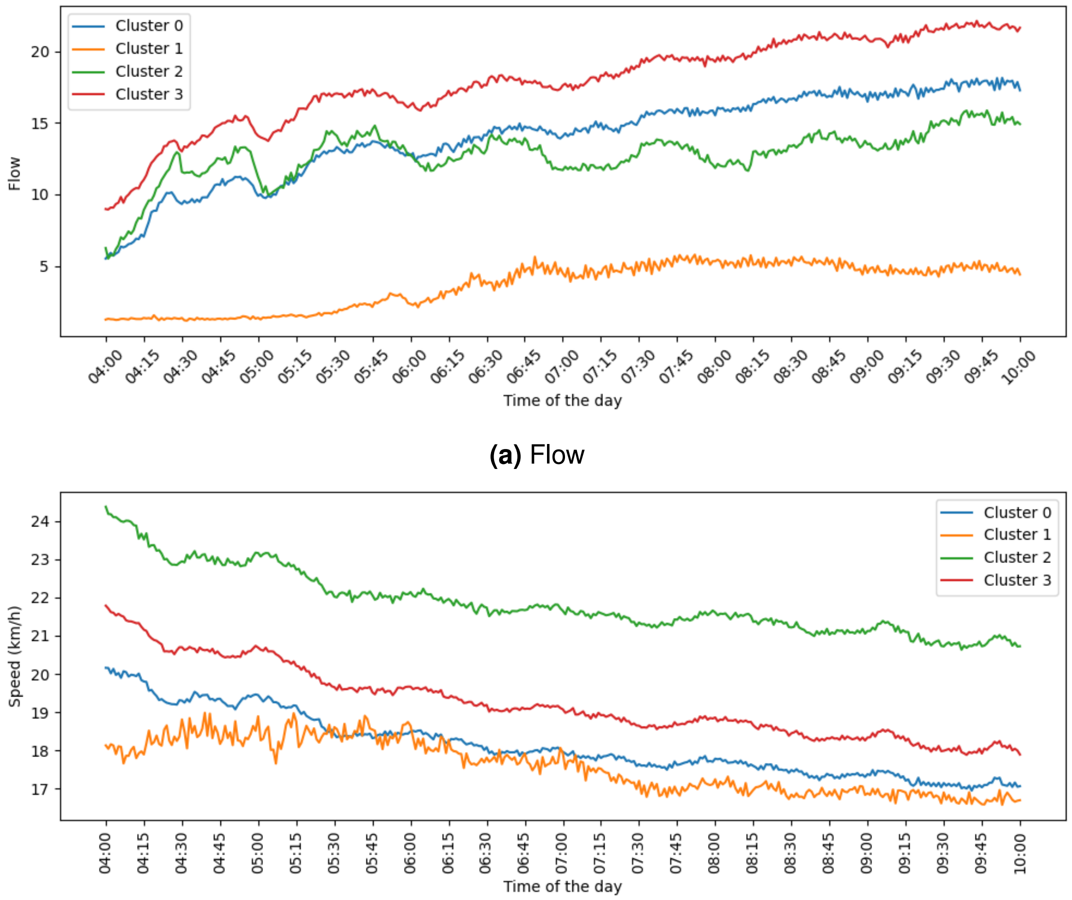
(b) Speed

Figure 2.3: Correlation between speed and flow in different sensors

Two very interesting finding can be found from this correlation matrix. First of all, it can be seen that it seems like there are two different groups of portals. Portal 55620, 56160, 56490, 56780 and 57055 are more similar to each other than to the other three, that also seem to form a group. Furthermore, speed and flow show different patterns. It can be seen that for the speed the correlation is generally higher within the same portal (indicated by high density of red squares close to the diagonal). On the other side, this is not necessarily the case for the flow values. It can be seen that the red squares are more spread out, showing that the correlation is also partially quite high with sensor data from other portals.

2.4. Clustering as lane identification

As part of the data analysis, lane identification was performed using unsupervised clustering. The underlying assumption is that different lanes exhibit distinct traffic characteristics in terms of speed and flow. Typically, faster lanes are located on the left (e.g., overtaking lanes), while slower lanes are positioned on the right (e.g., exit or merging lanes). It was hoped to find these behaviour patterns in the sensor data to infer lane structure. To uncover these latent lane groupings, K-Means clustering was used. It was chosen among several other clustering methods as it is the easiest method to control the number of clusters which should in the best case- represent the lanes. Based on the pattern that was observed in the correlation matrix showing two different portals groups, they were clustered separately. For the five portals, the number of clusters was set to 4 according to the number of lanes. Figure B.1 shows the typical dayprofil for the different clusters that were identified.



(a) Flow

(b) Speed

Figure 2.4: blabla

Based on these patterns it is very likely -though not proven - that cluster 2 corresponds to the leftest lane showing the highest speed, cluster 3 corresponds to the middle lane with a lower speed, cluster 0 the the right lane and cluster 1 to the lane where cars enter and leave the motorway with the lowest flow.

2.5. Data preprocessing

2.6. Further descriptive analysis for the portal 55620 and 56160

2.6.1. Correlation

2.6.2. title

3. Problem formulation

Formulate your speed or flow prediction problem and justify it by discussing its reasonableness for practical use in real-time application. For example,

What to predict? – Do you predict flow, speeds, or both? What is the problem type? – Formulate as a clustering problem, a regression problem, a classification problem, or a combination of these? What features to use? - Develop a-priori hypotheses about features you think are important in predicting 15, 30, 60 minutes in the future.

4. Methodology and evaluation methods

5. Model development

Data Preprocessing: Clean and preprocess the dataset if necessary, data normalization, and converting categorical variables into suitable formats for modeling. Feature Engineering: Analyze the dataset to identify relevant features that could impact bus arrival times. Create new features if necessary, such as time-based features, distance-related features, and aggregation of historical data. Exploratory Data Analysis (EDA): Perform EDA to gain insights into the relationships between different variables and bus arrival times. Visualize patterns, correlations, and outliers in the data. Model Selection: Select appropriate machine learning algorithms for short-term prediction task. Consider conventional machine learning models or deep neural network models depending on the nature of the data. Model Training: Split the dataset into training and validation sets. Train the selected models using the training data and fine-tune hyperparameters to achieve optimal performance. Model Evaluation: Evaluate the trained models using appropriate evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). Compare the performance of different models to select the best-performing one. For a comparative analysis, you need to develop at least three different types of models.

6. Model diagnostics

Design different scenarios to explore how the best model works overall. When and where it does not work? Is it robust to noisy and missing data in real-time? Is it generalizable over months?

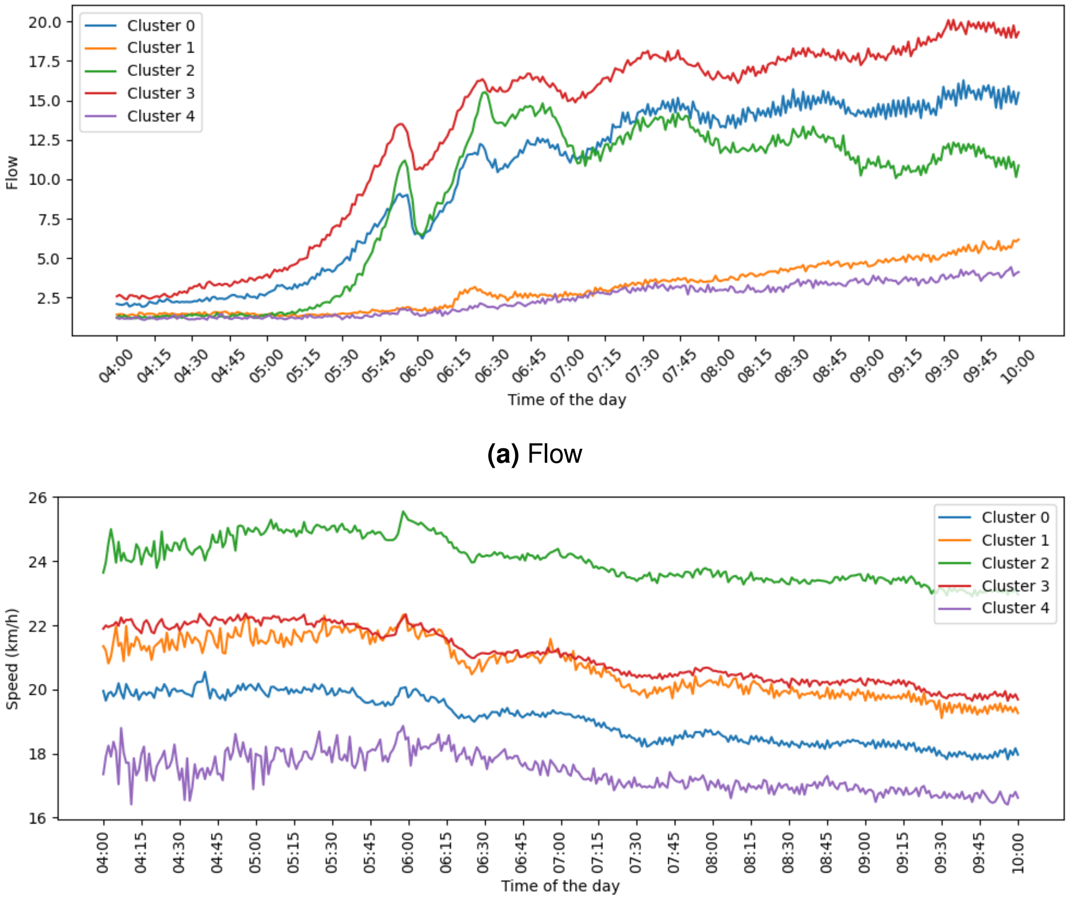
7. Model deployment

8. Summary of main findings and future works

A. GitHub code link

The project can be accessed under: `https://github.com/Jojo18-20/Project_AI_Transportation`

B. Additional graphs



(a) Flow

(b) Speed

Figure B.1: Clustering of the three other portals, the exact location of these sensors is not known, thus no interpretation is made on the cluster-lane