Data driven traffic management policy

Final report

Project: Datadriven policyutveckling

Authors:

Date: 20190528

ABSTRACT

Large amounts of data and information is generated for different purposes every day in the traffic system. There is an increased interest within the Swedish Transport Administration (STA) in using this data for planning of maintenance, traffic management and strategy work. In this report the first steps towards such a system is developed by the process of defining a business objective, collecting data, understand the data, prepare the data, create a model and evaluate the results. All these different steps were important in the performed study. To find a good case creating business value required lots of discussions and interviews with key figures at STA. The case investigated is to predict the traffic situation on four road segments in Gothenburg based on two years of data for the traffic situation, weather, and road situation including accidents and road works.

The data for primarily weather and traffic are not collected for the purpose of being used for this application. This is one reason for that data is missing from some of the data sets for different time periods. One conclusion from the project is that data analysts must be included not only in the data analyze phase, but also in the data collection phase to achieve good results.

Different methods for creating data driven models are evaluated and compared based on the two year period of data available. It is found that linear regression performs better than tree-based classification and prediction method regarding performance, while the tree-based method more clearly can create understanding for what variables that correlate to the traffic situation.

The methods for developing models based on the data used in this project are generic and are possible to be used when larger data sets are available. Additional data sources, such as events in the city and building works may also be included in such analysis. Furthermore, it is found valuable to have the possibility to develop models on a local computer based on a smaller data set, and make the final computations based on the larger data sets in a cloud based solution.

TABLE OF CONTENTS

TABLE OF FIGURES

1 PROJECT OBJECTIVES

In and around the traffic system in Sweden, large amounts of data and information are generated, from a variety of sources, which today are not utilized. The Swedish Transport Administration (STA) sees the potential of using this ecosystem of data for planning, maintenance, traffic management, follow-up, and strategy work. This project develops new methods for collecting and analyzing data, simulating/exploring policy options, using cloud infrastructure, and integrating data services and tools, and involving stakeholders in policy work.

The project is based on areas where new technology has the opportunity to promote development in line with the traffic policy goals for Sweden but where current policy needs to be developed to realize the latest technology. The idea behind the project is to use concrete method development - linked to passenger transport, freight traffic or professional traffic in the city - also take the first step to describe how traffic authorities at a general level can coordinate and experiment with data-driven policy development.

2 AREA FOR INVESTIGATION: PROACTIVE TRAFFIC INFORMATION

2.1 BACKGROUND

The project commenced in late 2017 and started by identifying areas that would be suitable for the project. The first user group feedback workshop took place in November, and the idea brought to the table at that time was based on a Machine Learning approach (Parker, Simari, Sliva, & Subrahmanian, 2014). This way, data can be used to drive policy development. The project then tried to extend the model into areas where it could improve policy-making within STA. Several compelling use cases (dangerous good transport, regular congestion, how to handle event-based congestion and traffic planning, high-capacity goods transport, snow clearance, road work/repair, and commuter traffic in the city) was identified. The model offered by (Parker, Simari, Sliva, & Subrahmanian, 2014) was based on the idea of mapping out the goals one wants to achieve and then matching these towards time series of data related to actions that have been taken and time series of data related to desired states. Unfortunately, the data necessary to perform such machine learning were either non-existing or were not possible to access for security reasons. The project thus needed to try a different approach.

In early 2018, the project embarked on addressing the challenges faced in the first part of the project. Among the project members, there was a consensus that the project should focus on the analysis of data and simulate/explore policy options. For this reason, a series of interviews were held with key stakeholders within STA, having expertise in both information security, data analysis, policy work, and digitalization strategy[1]. During these interviews, potential areas were discussed and assessed towards four essential requirements:

A. The prototype must provide STA business value
B. Necessary data must be both available and accessible
C. IBM Cloud should enable/lower the threshold for the prototype development

[1] See "Appendix 1: Interviews within STA" for more information on the interview findings

D. The prototype must be possible to develop within the project scope.

During these interviews several interesting areas were discussed, such as several use cases related to railways (such as predictive maintenance and report text mining), simulation of policy options connected to the introduction of autonomous vehicles as well as Mobility-as-a-service (MaaS).

These options were eventually dismissed due to not fulfilling the four requirements set out. The idea that the project finally settled at was proactive traffic information. Currently, traffic information is seldom provided proactively, but rather after the fact that a traffic incident has occurred. If STA could know in advance that tomorrow there will likely be severe traffic problems, this would affect how STA may inform travelers. The idea was assessed by the project group to meet all four requirements:

1. Proactive Traffic Information met a clear business need, expressed by TrafikGöteborg. Currently, they are evaluating short-time travel time prediction solutions and performs manual and more longer-term travel-time prediction. However, estimation on medium-term time ranges (such as the next day) was currently not implemented yet was found useful.
2. Data was available. There was a significant amount of passage data from the STA system STRESS, for running traffic, from 2014-01 to 2015-12.
3. IBM Cloud contained new capabilities that could be useful for prototype development.
4. The project development team assessed that a functional prototype could be developed within the time and budget constraints within the project.

As a next step, a focused study of factors influencing travel time was conducted.

2.2 Factors Influencing Travel Time

Travel time prediction is an import aspect of traffic management. For authorities, accurate predictions may serve as useful stimuli to perform necessary preparations. For travelers, accurate predictions allow them to either prepare for potential delay or choose another mode of transport to reach their destination faster.

As a basis, researchers typically use historical travel data (C. Schrader, Kornhauser, & M. Friese, 2004) (Domenichini, et al., 2012) (Yildirimoglu & Geroliminis, 2013) (Koesdwiady, Soua, & Karray, 2016) and then applies several variables to calculate each variable's impact. To correctly calculate travel times for a specific road segment or a network, there are two principal factors to consider (Lint, et al., 2004). First, travel demand, i.e., the number of vehicles that are predicted to utilize the road network. Second, the travel supply denotes the current capacity of the selected road network. Under proper conditions, travel supply may be close to equal of the network's maximum capacity, but travel supply may be significantly less during, e.g., heavy snowfall or accidents.

In the following, the variables found in the literature to have the most impact is presented.

2.2.1 Travel demand

When it comes to variables affecting travel demand, the most common parameter is temporal patterns (Yildirimoglu & Geroliminis, 2013), such as time of day and day of the week. For instance, during morning and evening commute, travel times often differ drastically compared to off-peak hours. Similar patterns can be observed when comparing work days and weekends. Other

temporal indicators include school holidays, general holidays, significant events, and even paydays.

In addition to temporal patterns, traffic information is observed to influence travel demand. For instance, if drivers are presented information from VMSs, radio broadcasts or navigation systems about severe delays or alternative routes (due to incidents), they are more prone to switch the mode of transport or choose ways that typically aren't used for the upcoming trip (Lint, et al., 2004).

2.2.2 Travel supply

Travel supply is dependent upon several variables. One such notable variable mentioned in the literature is *accidents* (Yildirimoglu & Geroliminis, 2013) (Domenichini, et al., 2012). Depending of the on the severity and magnitude of the accident, travel supply may decrease drastically or even be close to zero during accidents. Another related variable are *road works*, which are seen as mediators of travel supply.

Apart from such man-induced actions, *weather* is typically considered as an important variable (Hojati, Ferreira, & Charles). To what extent heavy rainfall affects road supply is disputed but there is broad consensus that snowfall do affect travel supply.

Finally, traffic control and road geometry are seen as an important variable. For instance, non-optimized traffic lights, ramp design and variable speed restrictions are considered to influence travel supply (Lint, et al., 2004).

2.3 SELECTION OF DATASETS

Based on this study, relevant datasets was sought after.

The base dataset came from STRESS1 traffic flow data for Göteborg between 2014-01 and 2015-12. The expectation is to extract valuable measure from the data that have two-way impacts to/from traffic management policy.

Given the focused study on factors influencing travel-time, the project in addition collected the following datasets:

Variable	Source	Accessibility	Comment
Temporal conditions	Calendars	Accessible	n/a
Accidents	STA	Accessible	This data is currently provided as open data through e.g. https://api.trafikinfo.trafikverket.se/ However, this source only covers current incidents. Therefore, STA made a separate dump of historical data between 2014-01 and 2015-12 from the STA system NTIS/TRISS.
Road works	STA	Accessible	
Road conditions	STA	Accessible	
Weather	SMHI	Accessible	We retrieved SMHI open data from https://opendata.smhi.se/

Variable	Source	Accessibility	Comment
Large events	City of Gothenburg	Accessible	This data is currently provided as a web site (https://nystart.trafikkontoret.goteborg.se/) However, the project needed more machine-readable data. To this end, the City of Gothenburg made a separate dump of historical data between 2014-01 and 2015-12 from the city's system "geodata".
Traffic information	Assessed out of scope	n/a	n/a
Road geometry	Assessed out of scope	n/a	n/a
Traffic control	Assessed out of scope	n/a	n/a

Table 1 - Datasets for factors influencing travel time

3 THE CRISP-DM FRAMEWORK

CRISP-DM (Wirth, 2000) (Cross-Industry Standard Process for Data Mining) is adopted in this project for the data mining approach. The CRISP-DM process model has been developed by a consortium of data mining users and suppliers including: DaimlerChrysler AG, SPSS, NCR, and OHRA.

As a methodology, it includes a descriptions of the typical phases of a project, the tasks involved with each phase, and an explanation of the relationships between these tasks. As a process model, CRISP-DM provides an overview of the data mining life cycle. The life cycle model consists of six phases with arrows indicating the most important and frequent dependencies between phases. The sequence of the phases is not strict. In fact, most projects move back and forth between phases as necessary. The CRISP-DM model is flexible and can be customized easily (IBM, 2011). Figure 1 illustrate the lifecycle as defined by CRISM-DM model.

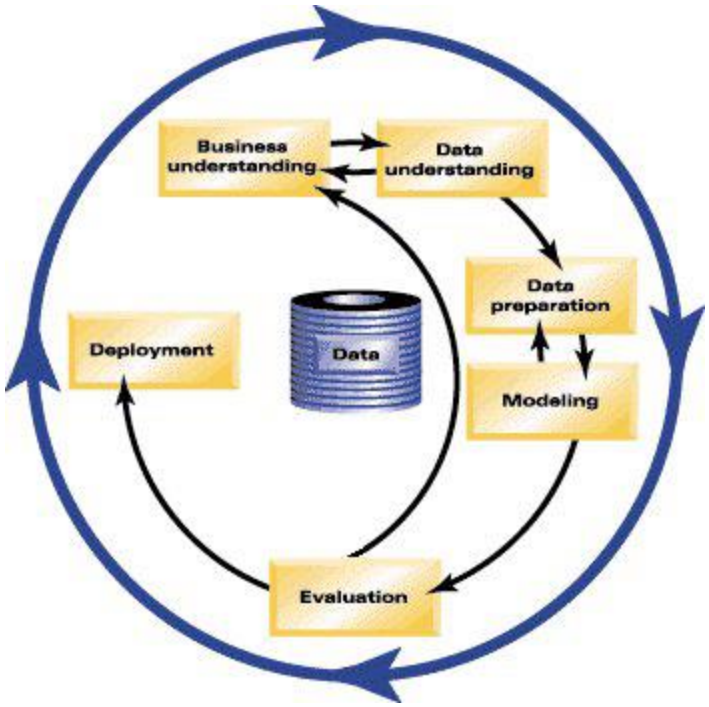

Figure 1: CRISP-DM lifecycle

4 BUSINESS UNDERSTANDING

The business understanding phase has been performed iteratively via interviews, data collection and understandings. The understandings of both business and available data are further refined after each iteration by different joined activities between the project team and different departments from STA.

4.1 DETERMINE BUSINESS OBJECTIVES

From the interviews (Appendix 1: Interviews within STA) and the accessibility to the datasets, we determine the business objectives as follows:

- Adoption of data driven approach in the long and short terms traffic management policies.
- Data preparation for all possible correlation analysis between different timeseries variables.
- Modeling traffic flows as functions of independent variables (such as weather, event and unwanted road situations)

Success criteria: Data readiness for analysis and a working proof of concept (prediction of tomorrow traffic situation).

4.2 ASSESS SITUATION

4.2.1 Resources

A number of datasets already exist today collecting real-time traffic related information. Datasets are maintained by different business departments. Roadwork planning data are mainly indicative

of the preferred time frame of related projects, thus it is difficult to acquire actual starting and ending timeline of a planned task. City event data are furthermore not captured with traffic management objectives. Policies are mainly subjective made by related personnel in compliance with guidelines and experiences.

4.2.2 Requirements, assumptions and constraints

The possible area of traffic management policy is inspired from the daily operation at TrafikGöteborg. The business summary is as follows; Every beginning of the week, TrafikGöteborg collect traffic related information from SMHI (weather data), planned roadwork, events, etc. and create a report as control guidelines for the next eight weeks. Difficulties found as the external road maintenance projects are planned individually with just time ranges, and are not very accurate in time.

From the limited available dataset (traffic flow), the project team plans to also collect data of the same time period for weather, city event and traffic situation (focusing on road works and accidents). The expectation is to identify possible correlations between these areas and find potential improvement to the traffic governance. The findings do not only focus on specific traffic control problems, but also to find potential correlations between information extracted from the available datasets and different traffic control policies. The adopted iterative approach helps to constantly update the business requirements in line with the availability and quality of collected datasets.

Very early findings are:

- Collected data has unfortunately too low quality to perform accurate time series analysis. No single sensor can provide a full 2 year of data, and data are missing at different time periods per different sensors. This creates unreliability when traffic flow information is aggregated from these sensors to represent e.g. a specific route segment.
- Seasonal impacts of traffic patterns can hardly be extracted because of inadequacy of data. The input datasets contain 2 years of traffic data. Due to data missing at different time periods, there is rarely the case where the same period of the year occurring more than once in the training datasets.
- Therefore, we decide to focus more on daily traffic flow pattern instead: Traffic patterns are found to be different between working and nonworking days (see Figure 6) in a nonlinear relation to hour of day.
- The sampling frequency and the durations of different datasets are different. Large number of events with long duration and inaccurate start/end time may mislead the impact analysis.

The data mining models are thus only chosen from classification and forecasting classes. The resulting algorithms from the models should also be simple and high clarity. Experiments should be repeatable and prepared to easy include new datasets or new techniques upon data quality levels.

4.2.3 Terminology

The following terminologies are used throughout the project:

No	Terms	Description	Unit

1	Traffic volume	Number of vehicles passing the detection zone during a short data collection time period	#car/minute
2	Occupancy	Time occupancy calculated as sum of durations of all vehicles traversing the detection zone in the data collection time period and dividing the total duration by the time period	%
3	Average speed	Average speed of vehicles passing the detection zone during a short data collection time period	km/h
4	Density	Number of vehicles per km per lane at a given time	#car/km
5	Event	Göteborg city events registered with municipality	
6	Roadwork	Part of the road, or in rare cases, the entire road, has to be occupied for work relating to the road, most often in the case of road surface repairs.	

4.3 DETERMINE DATA MINING GOALS

In this step, we identify the business areas where data can assist policy makers, together with potential data quality requirements to improve the data capture process for the next steps. The latter is critical for any adopter of data driven approaches, though often being underestimated in many data mining projects.

- Classes of data mining models:
    - Classification: Different models with binary outcomes (yes or no). This family of models include: Neural Net, C&R Tree, QUEST, CHAID, C5.0, Logistic Regression, Decision List, Bayes Net, Discriminant, Nearest Neighbor, SVM…
    - Clustering: The class of models identifying groups of records that have similar characteristics, including: TwoStep, K-Means, and Kohonen
    - **Prediction:** The class of models with numeric range outcomes. Supported algorithms include neural networks, C&R Tree, CHAID, linear regression, generalized linear regression, and support vector machines (SVM). The models can be compared based on correlation, relative error, or number of variables used. Different combinations of stepwise, forward, and backward regression methods can also be used.
- Standardization of time interval: We decide to reconcile the different dataset to the same timescale with either 1 hour or 5 minute intervals.
- All the analysis should be done with datetime information. Datasets are linked via timestamp and geospatial information.
- Find the best model(s) and model parameters upon the data quality level, i.e. no fixed desired outcomes are defined. We assume that the same approaches can be applied and provide improved results when data management policy is in place to monitor and improve data quality.

5 DATA UNDERSTANDING

5.1 COLLECT INITIAL DATA

5.1.1 Traffic flow datasets

Monthly traffic flow data in forms of CSV files are provided from STRESS1 system. These data files are uploaded into a single PostgresSQL database table (partitioned by date and/or detector ID). The data is collected from 254 traffic sensors in Göteborg. The locations of these sensors are illustrated in Figure 2, covering several national roads. The interval between two consecutive measures is approximately one minute. Each sensor provides per a one minute measured period the following information: speed, occupancy, head way and number of vehicles per lane where the sensor located

Figure 2: STRESS1 sensor locations in Göteborg

The sensor descriptions are merged, using sensor ID, from the original provided sensor specification file and later from the STRESS2 sensor specification provided by STA. STRESS2 sensor specification provides a more complete and accurate description of the sensors (e.g. including also Lane ID and reconciled geocoordinates of sensors of the same group). However, some sensors available in STRESS1 do not exist in the STRESS2 sensor specification data, thus these sensors will have incomplete specification information.

5.1.2 Weather datasets

Weather data are first collected from SMHI open data API and then from STA. SMHI datasets for 2014-2015 are not complete, and at a later phase of the project we decide to only use the dataset provided by STA using VViS. The weather dataset provides temperatures (Road, Air, Dew point), humidity, wind, severe conditions (Nedbtype and level of rain, snow, melted in mm). Time intervals between two consecutive measures are approx. 30 minutes. The data is aggregated from the SMHI weather stations in Göteborg and from STA weather sensors.

5.1.3 Traffic situation datasets

The dataset includes traffic situations such as accidents and roadworks. The datasets are provided in 3 separate CSV files denoted:

- Väglagsstatus
- Objektversioner
- Väglag

5.1.4 City events

The dataset provided as single csv file logging all registered events in Göteborg for 2014-2015.

5.1.5 Sensor specification

As mentioned earlier sensor specification is aggregated from two provided dataset:

- Detektorpositioner.zip: STRESS1 spec
- Detectornames.csv: STRESS2 spec

5.2 DESCRIBE DATA

Data file format: CSV files with timestamp (different sampling) and geo-locations. Data size: 125GB DB size.

5.2.1 Traffic flow data specification

The CSV files are collected by month: gbgData_<yyyyMM>.zip

Provided description of fields are:

- [DetectorID] - Detektoridentitet
- [VehicleClassID] - fordonsklass
- [Timestamp] - När mätningen gjordes
- [Flow] - Flöde per timme
- [Speed] - hastighet
- [Occupancy] - Beläggning
- [Confidence] - Konfidensvärde
- [Tdiff] - Tidsdifferens hämtning/mätning
- [Timecycle] - Mätperiod i minuter
- [NoVehicles] - Antal fordon per aggregeringsperiod
- [Headway] - Fordonlucka
- [MeasuresIncluded] - Andel mätningar som ingår i aggregeringintervallet: 1 = 100%

VehicleClassID values are interpreted as follows:

- 0: Alla fordon/Ingen fordonsklass angiven
- 1: Personbil
- 2: Personbil med släp
- 3: Lastbil
- 4: Lastbil med släp
- 5: Buss
- 6: Övriga fordon

5.2.2 Weather data specification

Subsytem: VViS mätdatauttag

Kördatum: 2018-11-02 13:21

Tidsperiod from: 2014-01-01 00:00 till 2015-12-31 23:31

Mätstation: 1423, Högen

Available attributes:

- Tidpunkt: Timestamp yyyy-MM-dd HH:mm
- TLuft °C: Air temperature
- TYta °C: Road temperature
- Daggp °C: Dew point
- Lufu %: Humidity
- Vind m/s: Wind
- Vindmax m/s: Wind max
- Virik °: Win direction
- Nedbtyp: Severe weather type
- Snö mm: Level of snow
- Regn mm: Level of rain
- Smält mm: Level of melted

5.2.3 Event data specification

Event dataset is provided as an XLS file from the Database query

- EventName
- Location
- CustomerName
- Contact
- ContactPhone
- ExpectedAudience
- PoliceDNR
- PriorityName
- Priority
- Color
- PreEstablishmentFromDate

- PostEstablishmentToDate
- EventFromDate
- EventToDate
- FromDate
- ToDate
- WKT_SWEREF_99_12_00

5.2.4 Road situation data specification

5.2.4.1 Väglagsstatus

CSV file with following header fields:

Län, Starttid, Sluttid, obhistid, verhistid, Uppdateringstid, väglagsstatustext

5.2.4.2 objektversioner

CSV file with following header fields:

objekttypskod,objekttyp,objektID,skapandetidpunkt,arkiveringstidpunkt,versionsID,versionslopnummer,versionens_starttidpunkt,versionens_starttidpunkt_ar,versionens_starttidpunkt_manad,versionens_starttidpunkt_dag,versionens_starttidpunkt_veckodag,versionens_starttidpunkt_timme,versionens_starttidpunkt_halvtimme,versionens_starttidpunkt_kvart,versionens_starttidpunkt_minut,versionens_stopptidpunkt,rubrik,beraknad_starttidpunkt,beraknad_stopptidpunkt,DiffStarttidUppdateringstid_mi,paverkanskod,Paverkan,statusens_starttidpunkt,statusens_stopptidpunkt,statuskod,status,x,y,lagestext,Lan,primart_vagdata,vagnamn,vagnummer,handelsekod,handelsetext,tillaggstext,tillf_begr_typ,tillf_begr_varde,beskrivning,franplats,tillplats,paverkad_trafikriktning,vag_avstangd

5.2.4.3 väglag

CSV file with following header fields:

objekttypskod,objekttyp,objektID,Regionnr,versionsID,versionslöpnummer,versionens_starttid,versionens_stopptid,varaktighet_mi,x,y,länsnr,vägnummer,frånort,tillort,väglagssträckenummer,väglagssträcka,väglagskod,väglag1,väglag2,varningskod,varning

5.3 EXPLORE DATA

The datasets are then uploaded into PostgresSQL database tables with selected attributes of interest together with geometry information. The first exploration iteration of different datasets have been performed to achieve required insights into the data and possible patterns, possible data mappings, etc.

5.3.1 Spatial data mapping

The datasets are exported from different legacy system, thus use different coordinate systems. We use the QGIS tool (QGIS Development Team, 2019) and the PostGIS extension of PostgresSQL to provide the spatial insights of the data. The geocoordinates are standardized into the EPSG4326 spatial reference. Figure 3 shows locations of the 254 sensors that provide the traffic flow data used in the project. This provides the insights regarding which route segments that could be of

interest to study, and also enable visual inspection of the relationship between sensors in the traffic flow network. The right panel shows available information of a selected sensor as extracted from the dataset.

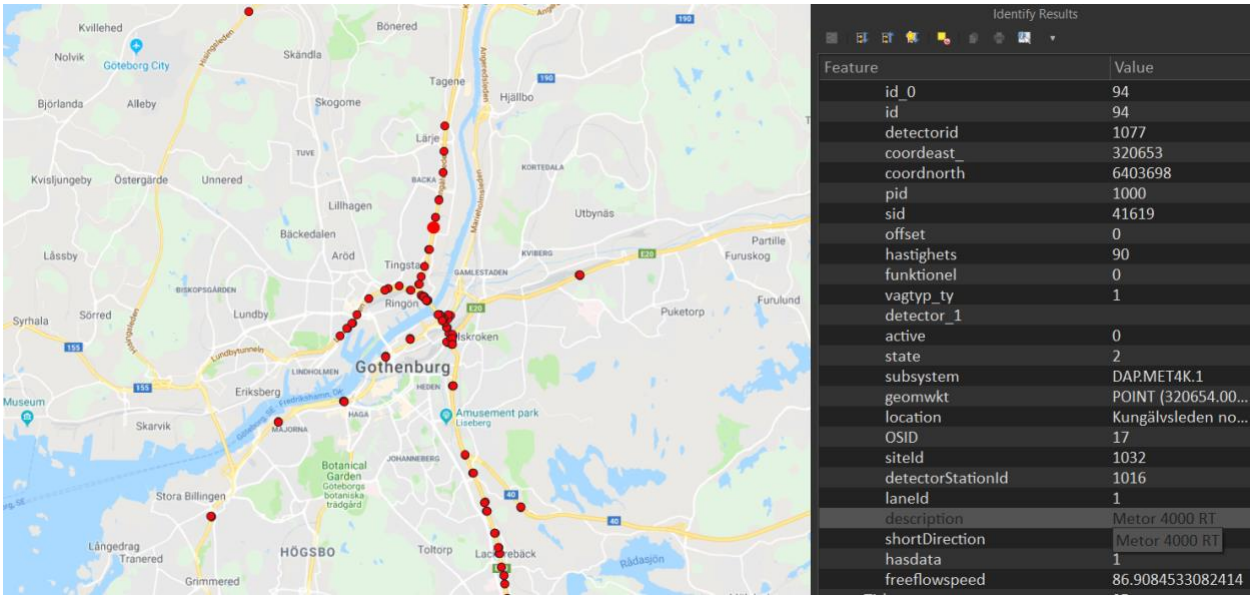

Figure 3: Sensor location data

The event data spatial distribution is illustrated in Figure 4. The events are not located on route segments, and thus do not have a clear relationship to the route segment structure.

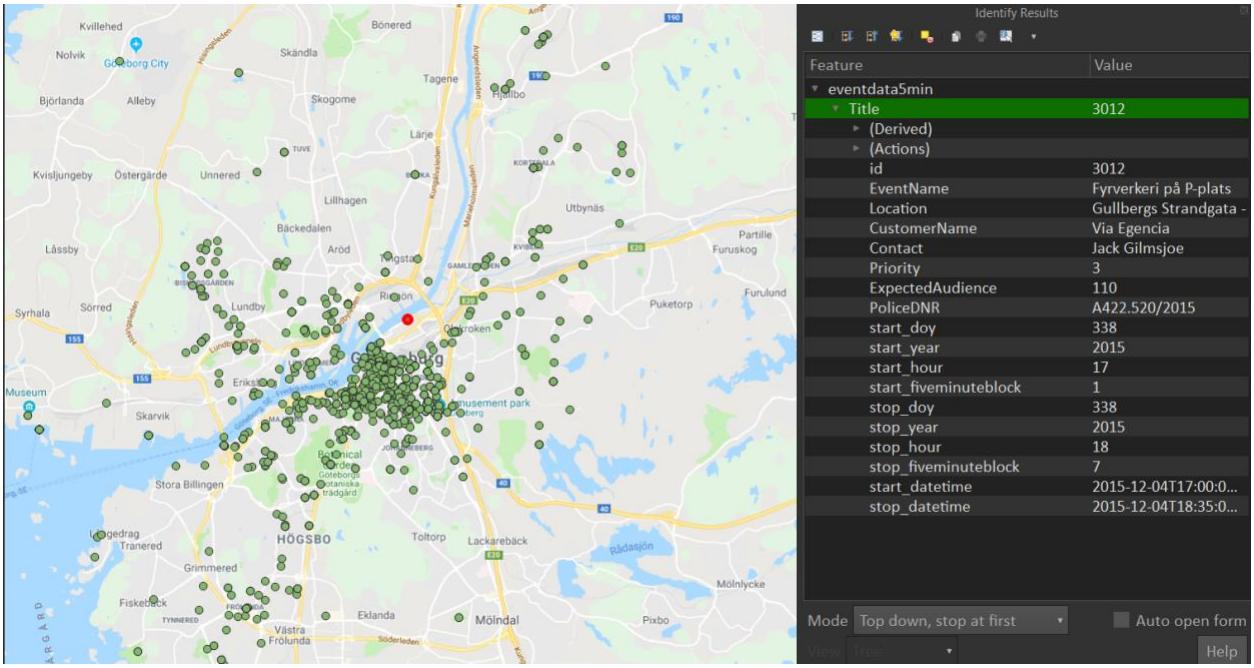

Figure 4: Event data

The road situations are located on route segments (see Figure 5). However, the dataset provides situations of the greater Göteborg region, and involve a lot more route segments where there is no sensor located (comparing with sensor distribution in Figure 3).

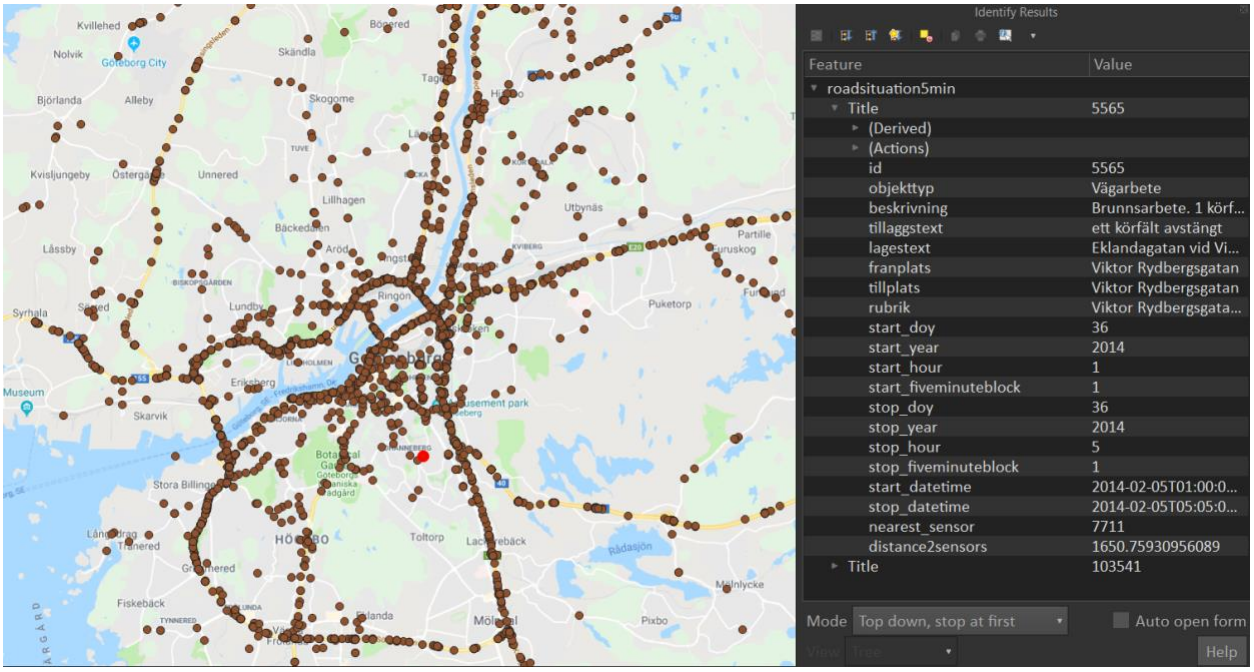

Figure 5: Road situation data

5.3.2 Traffic flow data

The traffic flow data is explored by plotting speed, volume and occupancy against timeline. The first observation is that there are two main different traffic patterns that describe typical working days and non-working days, i.e. weekend and Sweden public holidays.

We then explore these patterns as illustrated in Figure 6. As can be observed from the figure, speed drops are found at two peak hours during working days (8:00am and 4:00pm). The speed decrease during peak hours are more noticeable in Kungsbackaleden North (Kn) than in Kungsbackaleden South (Ks).

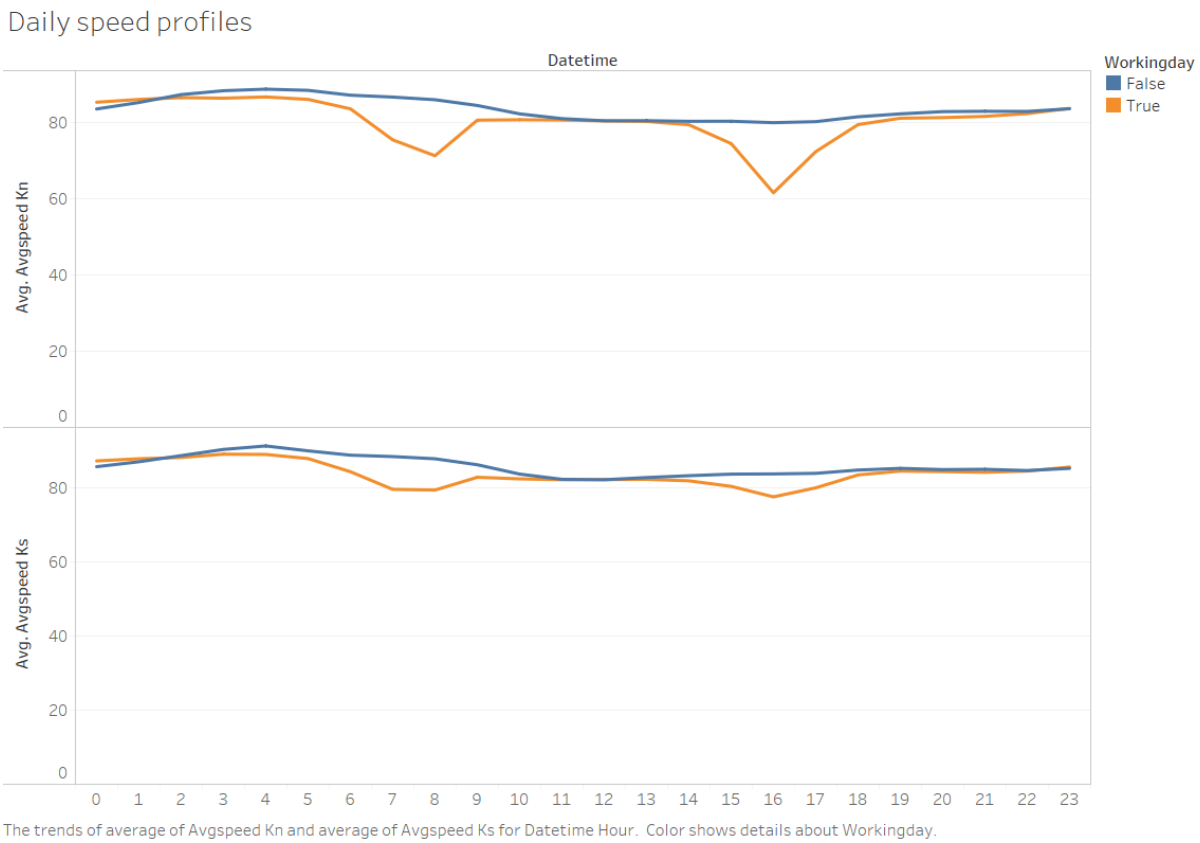

Figure 6: Average speed working vs non working day

Figure 7 shows how speeds at different time of the day are impacted by the road situation. Besides the two peak hours, we observe that speeds are strongly impacted by accidents happening at 5:00am at Kungsbackaleden in both directions.

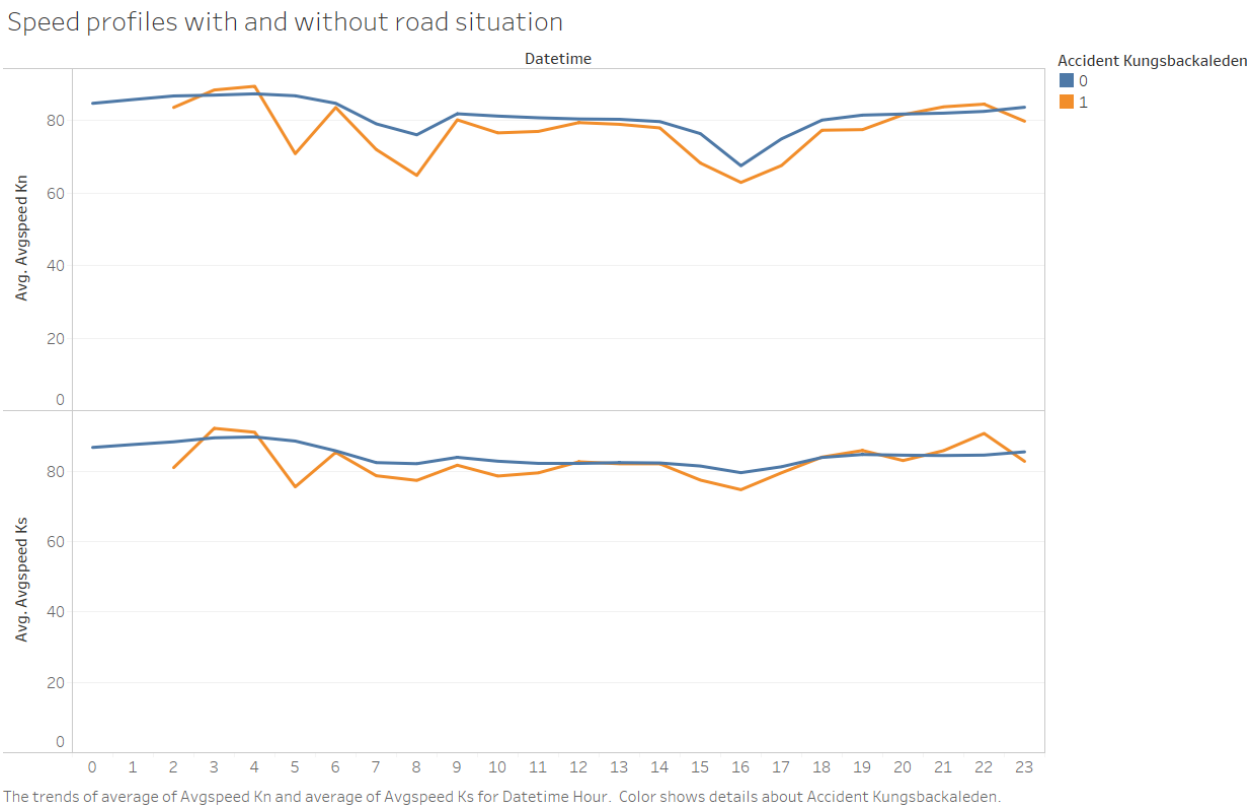

Figure 7: Average speed Kungsbackaleden with and without roadwork/accident

In Figure 8, we observe that the road situation does not cause much impact to the traffic load (volume and occupancy) both in working and nonworking days.

The standard deviation distribution of traffic flow ratio, which is the ratio between the travelling speed and the free flow speed, over the hours of the day is shown in Figure 9. We observe that during the peak hours of the working days, the speeds vary much more than other times, which will result in less accurate in estimations of speeds in peak hours comparing to other time slots.

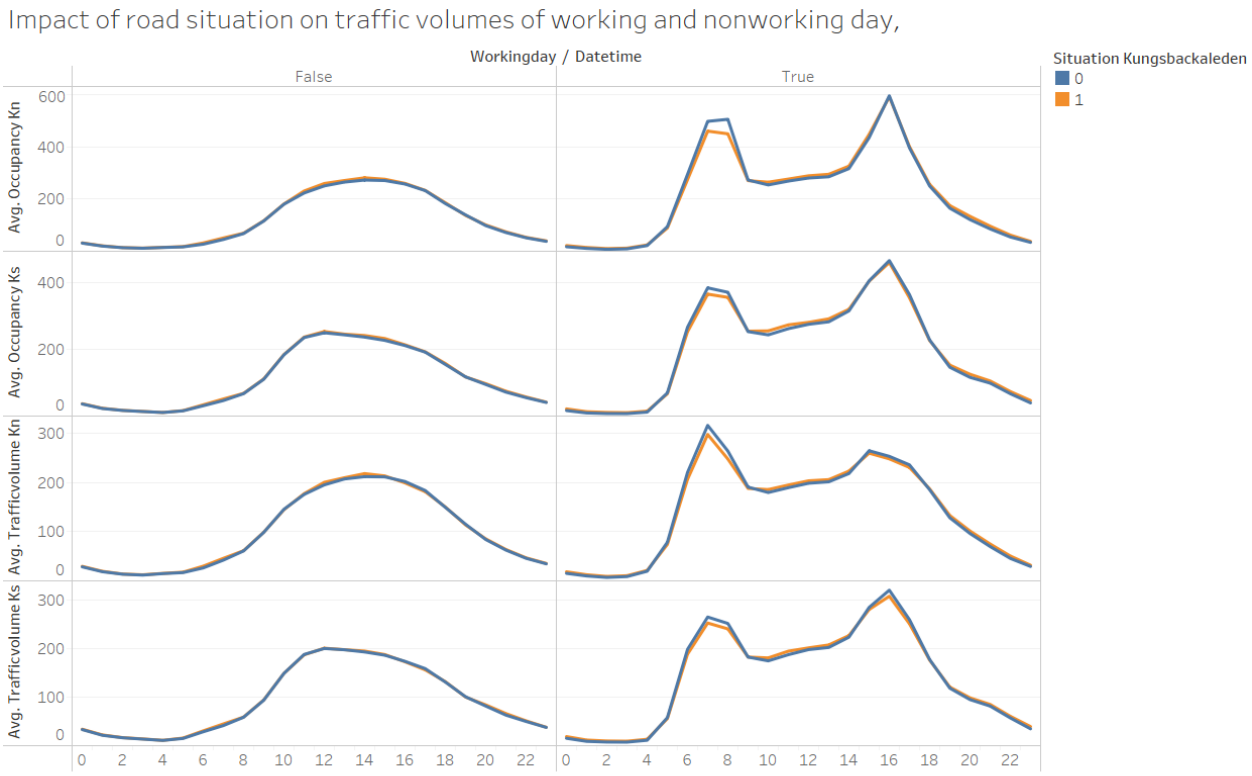

Figure 8: Daily traffic volume of Kungsbackaleden with and without traffic situation

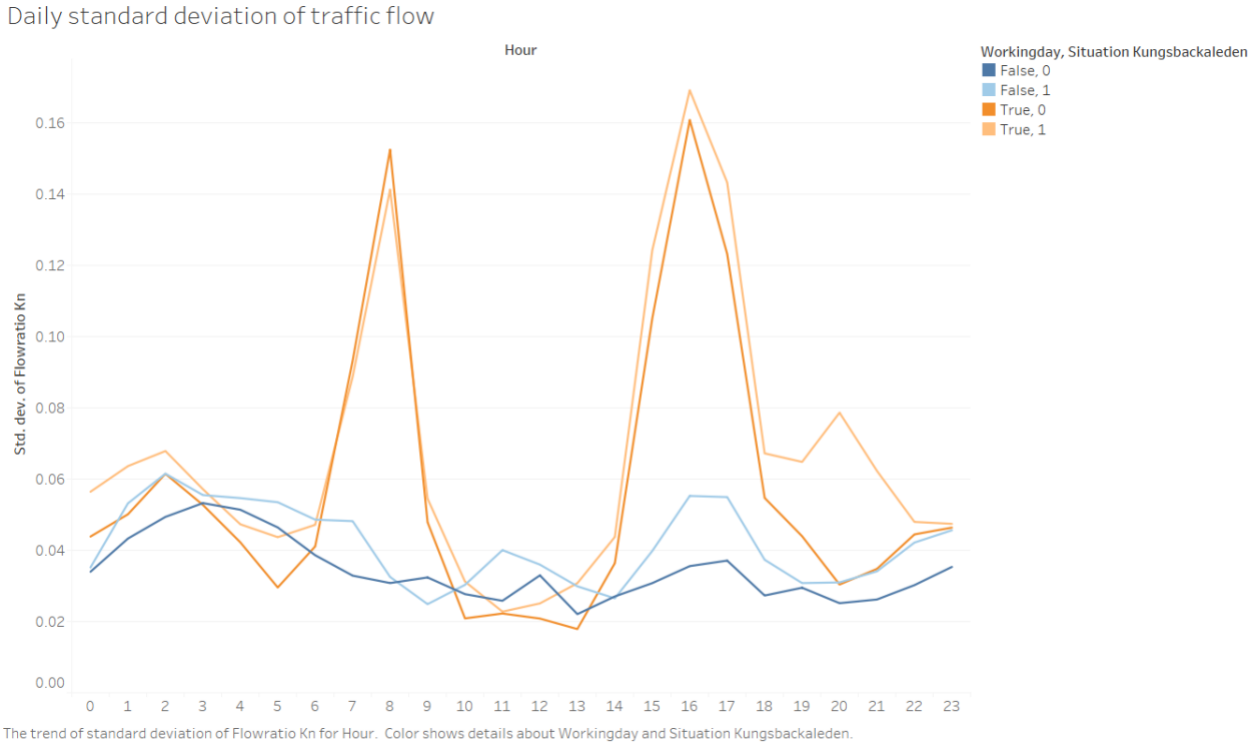

Figure 9: Standard deviation of traffic flow per hour of day

5.3.3 Weather data

The weather data has several continuous measures including temperatures (air, road and dew point), humidity percentage, and wind speed. It also contains severe weather condition (rain, snow) with related parameters. Except for severe weather conditions, weather data is usually quite smooth during the day. Figure 10 illustrates the temperature distribution, and Figure 11 illustrates snow and rain level distribution over the timescale of the year. Air temperature data has not been recorded in the dataset during the summer months of 2015.

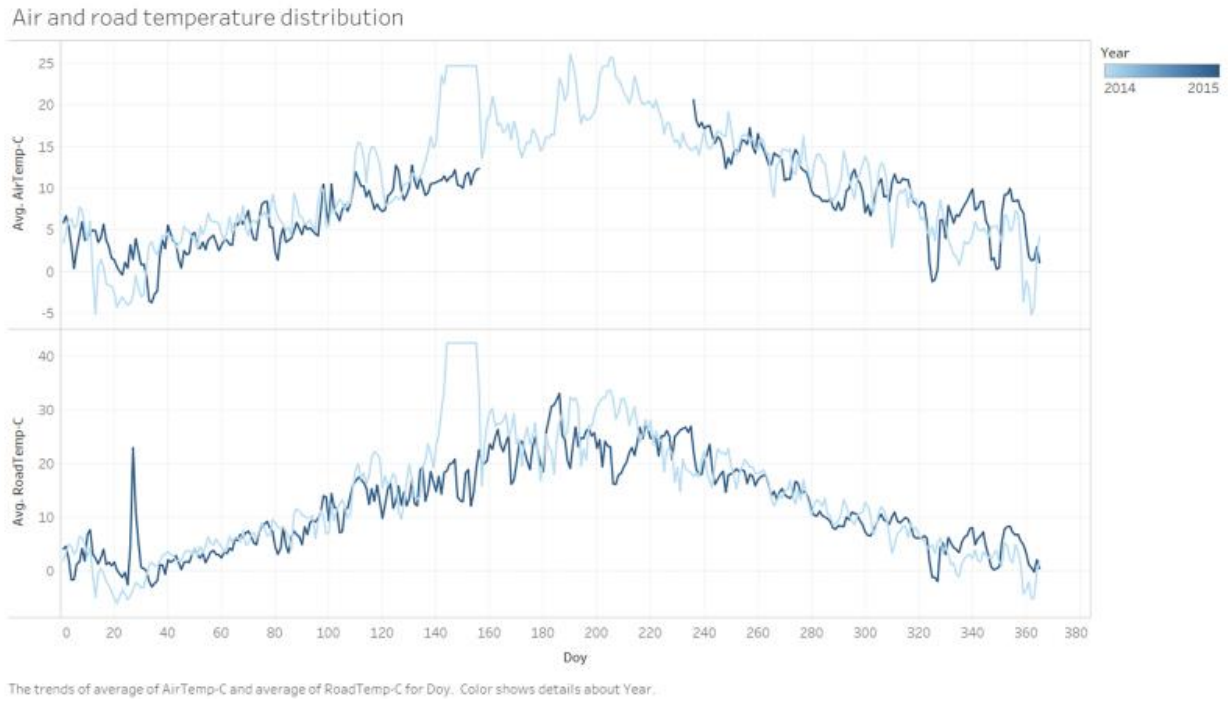

Figure 10: Temperature data over 2014-2015

Figure 12 show in average, how different severe weather conditions create impacts on traffic loads at the Kungsbackaleden route segment. The impacts are different per different time of the day depending on if the day is a working day or not. During nonworking days, severe weather condition leads to a decrease in traffic loads (volume and occupancy), while that is not the case in working days. Occupancies during peak hours of working days tend to increase when there is a raining condition. This can also be observed from Figure 13, where traffic flows are impacted most by rainy condition during working day peak hours, resulting in higher road occupancy.

During nonworking days, the largest speed reductions are observed when there is snow.

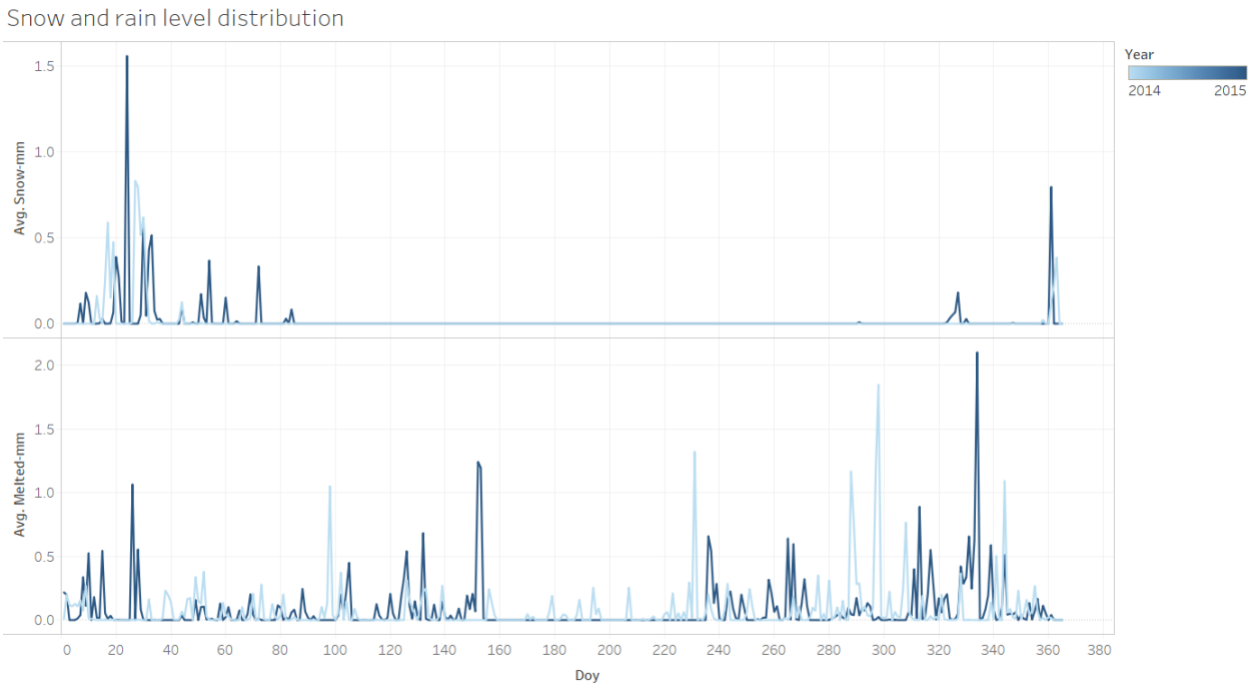

Figure 11: Snow and melted level 2014-2015

Figure 12: Traffic volume daily patterns of different severe condition of weather

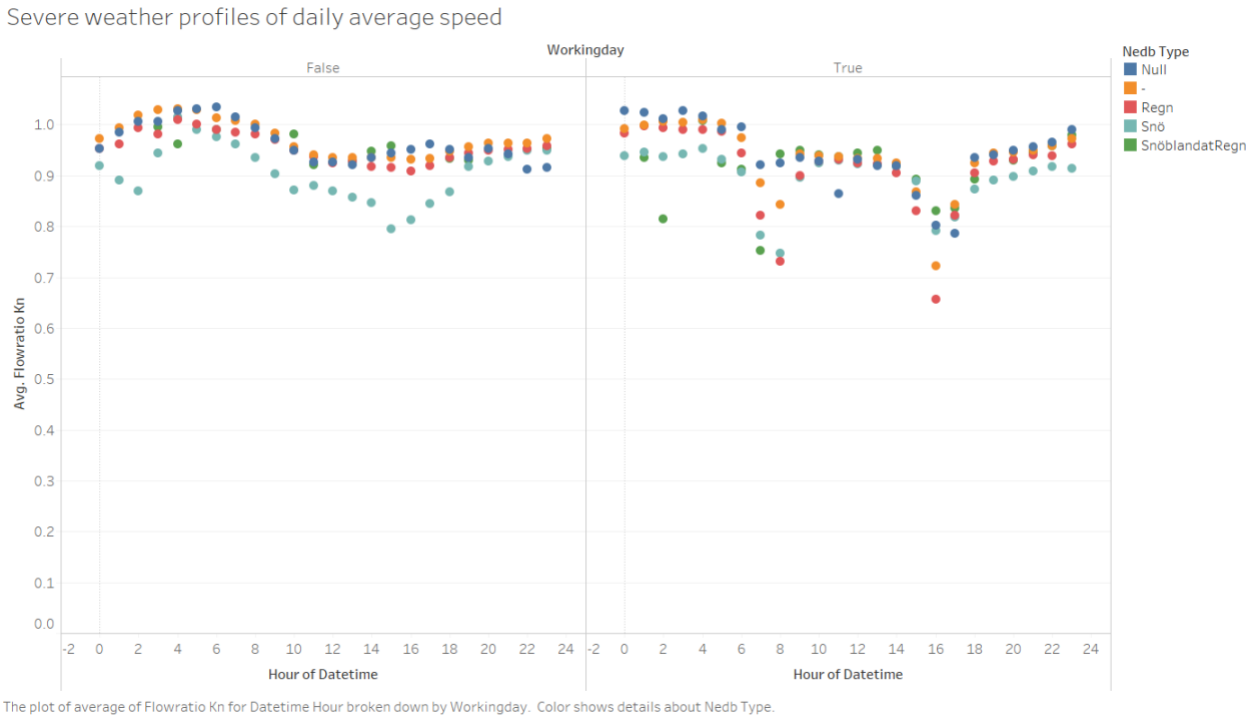

Figure 13: Impact of severe weather condition on daily traffic flow (Kungsbackaleden)

5.3.4 Road situation data

Road situation data includes road works and accidents. Figure 14 and Figure 15 show the relation between the number of accidents and traffic speed and volume. The accident distributions actually follow the normal distribution of traffic speeds and volumes. This means that there is no clear relationship between traffic flow to probability of accidents if we only investigate from the current datasets. More independent variables should be included to further investigate which parameter will increase the accident risk.

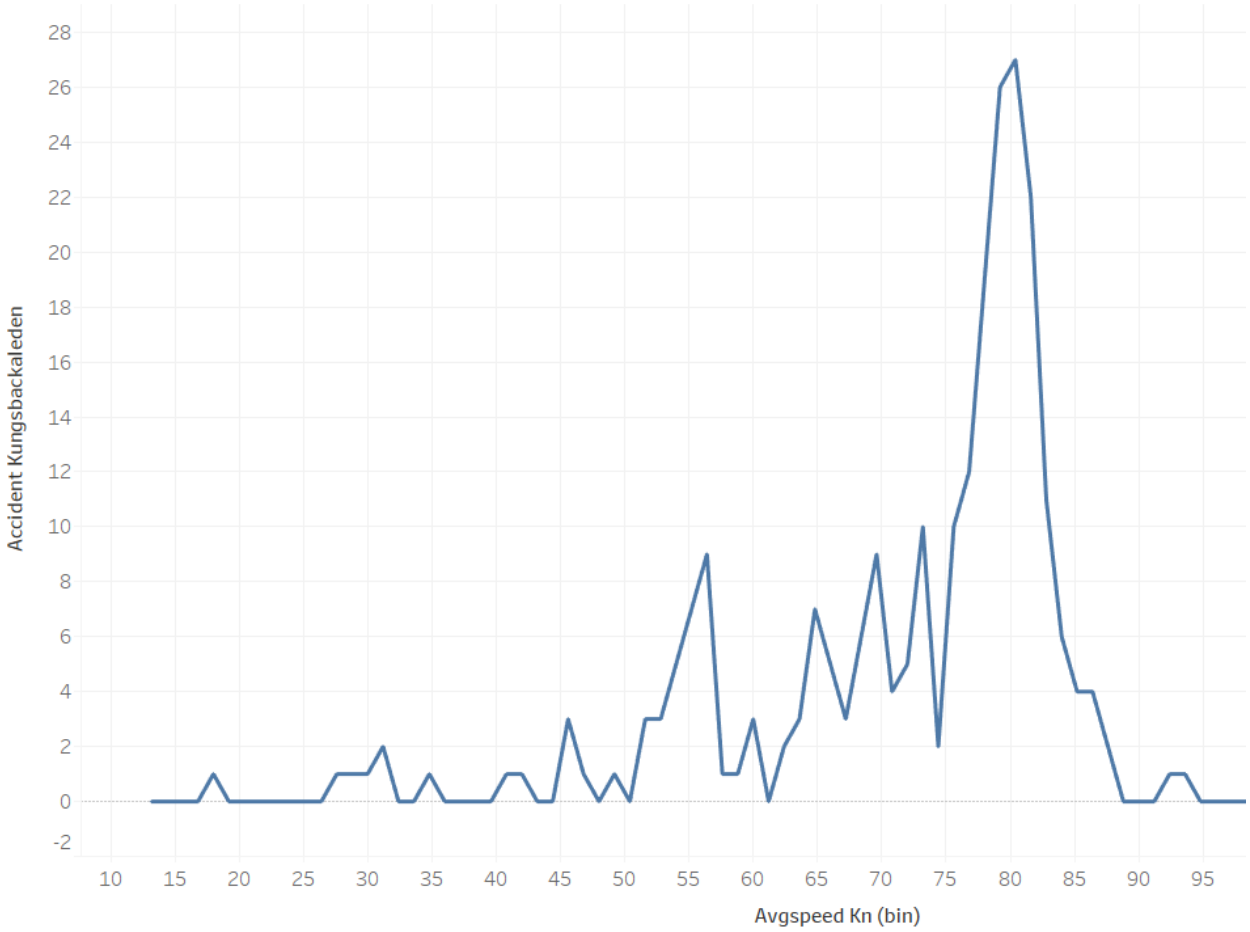

Figure 14: Distribution of accidents at different speed ranges in Kungsbackaleden

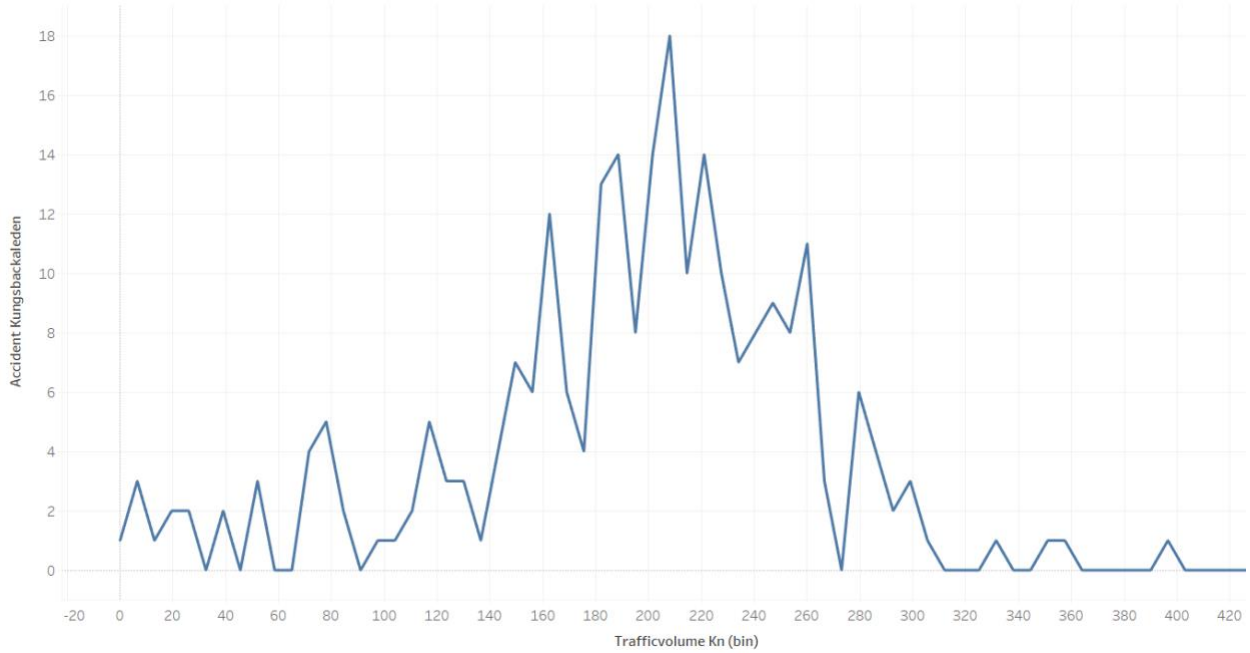

Figure 15: Distribution of accidents at different traffic volume ranges in Kungsbackaleden

5.4 VERIFY DATA QUALITY

We verify the data quality of different datasets with the following criteria:

- Traffic flow datasets:
    - Completeness of data over the timeline (possibly use durations where there is no missing data holes).
    - Completeness of data over spatial area by investigating if all selected sensors are providing data. Find outliers such as out of range values, e.g. speed over 200km/h.
    - Sensor signals having constant values over long period.
- Event and road situation data:
    - Location quality: Possibility to link between an event or situation to a sensor or group of sensors based on route network structure.
    - Timeline quality: Accuracy of start and end time. Duration and co-occurences.
    - Impact level measures: Audience size of event, or lane ID which is closed following a traffic situation.
    - Time resolution of an event or situation.
- Weather data:
    - Consistency between weather stations.
    - Completeness of weather attributes over time.

Followings big data quality issues are found:

- Each timeseries data has different data qualities over the timescale.
- Incompleteness of short-terms (e.g. one day) historical data
    - 477/730 days where there exist sensors providing some data every hour during one day for the 4 selected best road segments.
    - 443/730 days where the previous day also have 24 hour data.
- Unreliable of the sensor data (e.g. speed of over 200km/h or defaulted value). Speed of values greater than 249 are actually used as error codes in STRESS1 system. As seen in Figure 16, there is no clear separation between error codes (250+) and measured speeds. This results in difficulty in outlier detection rules.
- Road situation data are not in searchable category, i.e. there is no structural link to a route segment.
- Event data does not provide accurate start time and end time. Event durations are indicative information and often cross-day. Since the speed variation within the day is larger than intra days, cross day events are not useful in the prediction of the speed.

The data from the STRESS system is plotted for different road segments to understand the different aspects of the data. Figure 16 presents the speed for all nine sensors available in Tingstadstunneln for all time samples during this two year period. As can be seen, the different aspects mentioned above can be found in this figure, and is a basis for the selection of the data sets to be used in the analysis in the latter parts of this report.

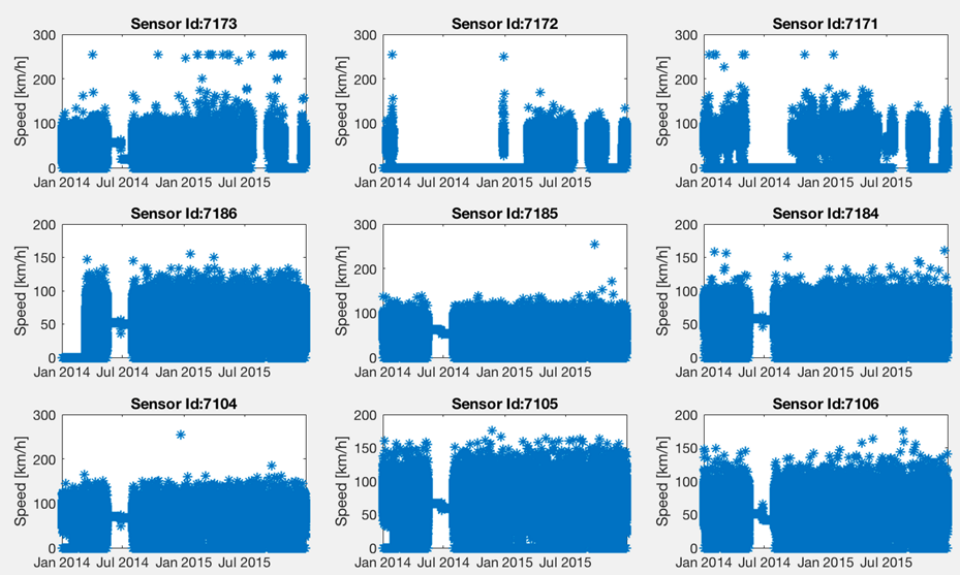

Figure 16: Data availability of different sensors in the selected segments

6 DATA PREPARATION

6.1 SELECT DATA

6.1.1 Traffic flow data

Traffic flow data are provided as csv files per month for all available sensors during the 2 years period 2014-2015. The sampling rate is approximately one minute. The flow data per sensor is classified also by car types (car, truck, bus…). Without losing the generality of the experiments, we decide to not split the experiment per car type, thus only vehicle type of 0 (all) is used in the next steps.

The traffic related information are selected as: traffic volume (number of cars per measured period), occupancy (time slot occupied by car), and speed. Since the interest focuses on traffic flow, in most of the experiments, the speeds are used as the main measures.

For historical data related analyses, only records that have one-day complete measures are considered in selection of training data. Where applicable, we only select records that are complete (i.e. having information in all required fields). Finally, the traffic flow data are aggregated into timescales of both 5 minute intervals and 1 hour intervals.

6.1.2 Weather data

Weather data was first extracted from SMHI Open data (https://www.smhi.se/klimatdata/utforskaren-oppna-data/). However we decided to use the weather data provided by STA with more traffic related accumulated information and better timescale resolution.

To achieve the same time interval as for the traffic flow data, we replicate weather data to create a higher resolution of 5 minute interval timescale from the original 30 minute interval scale.

6.1.3 Road situation data

From the provided data, we only select road situation data that are of categories roadworks or accidents witch are located in the polygons (manually selected using QGIS tool) that cover the selected road segments of interest (see Figure 17)

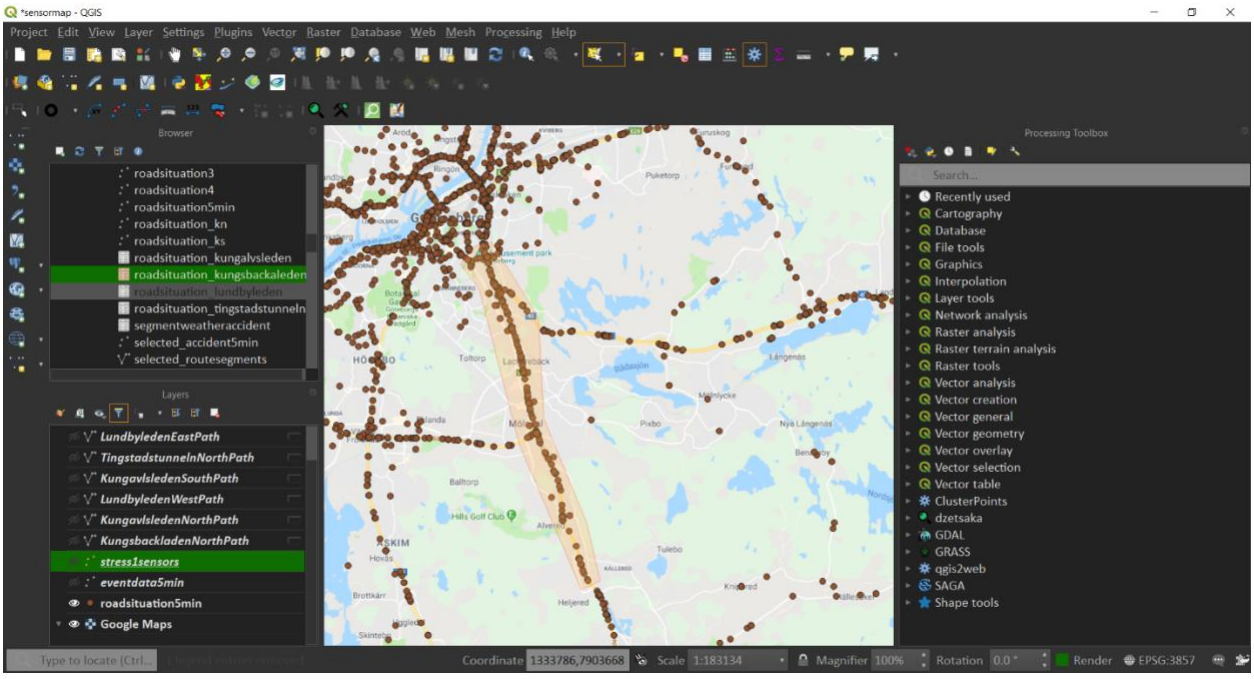

Figure 17: Selection of road situation related to a route segment

6.1.4 City event data

City event data is extracted from Göteborg Open data. The data provides event schedules and geolocations of registered events in Göteborg during 2014-2015 (see Figure 19). Since the quality of time information in the event dataset is not enough for impact analysis, we decide not to use event data for next steps.

Instead, the investigation of the event data can potentially be performed manually in reverse order. We start with some observed issues in the traffic flow, and investigate the potential cause from the event data by time correlation. This may result in the relation to start time, end time, or some other information.

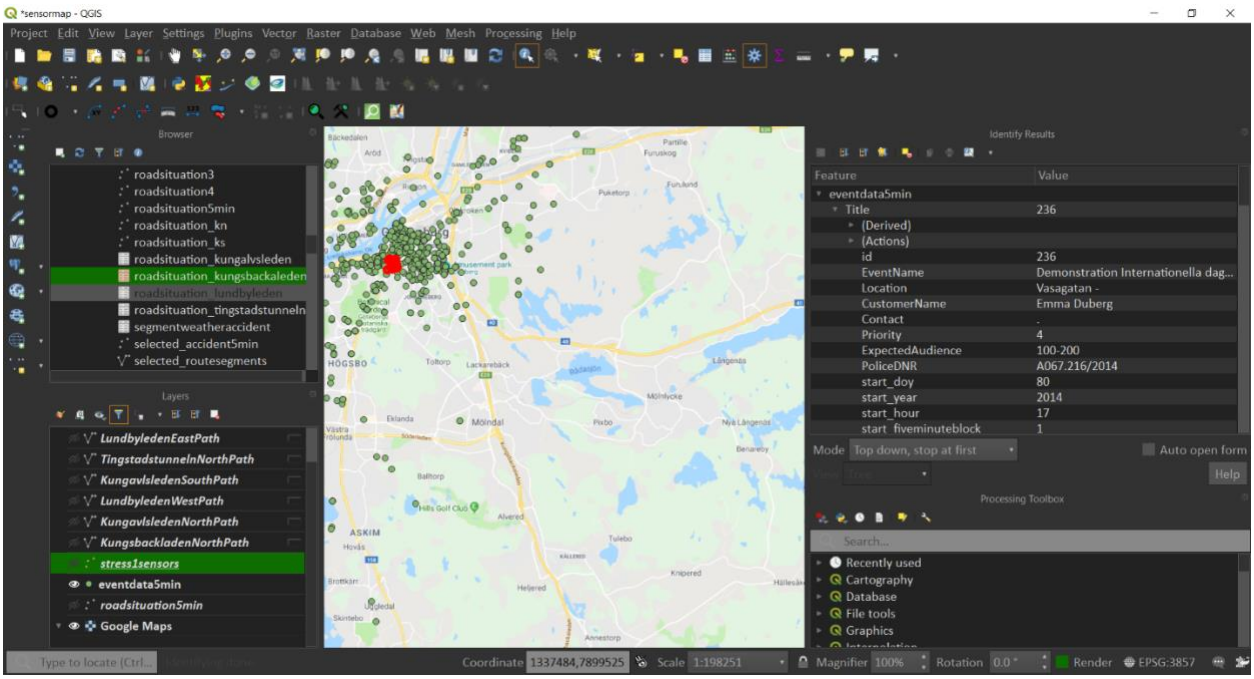

Figure 19: City event in Göteborg 2014-2015

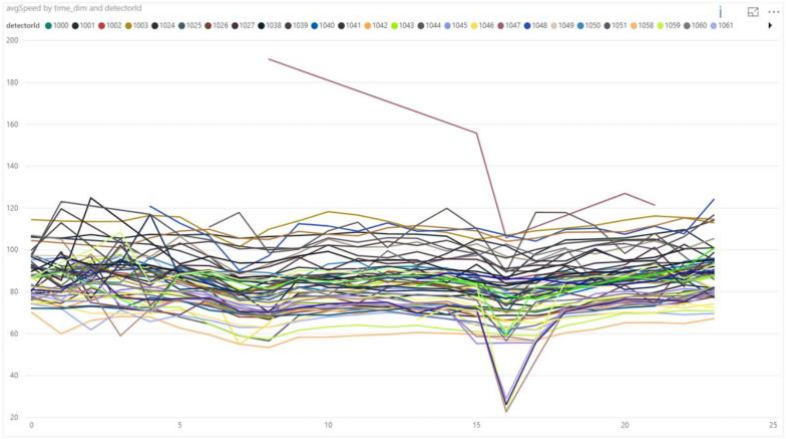

Figure 18: Average speed for different sensors 2014-01-08. The day is a weekend day

An example is the investigation of the drop of speed observed by a number of sensors by 4:00pm 2014-01-08. The related event as found in the event dataset is as below:

id	EventName	Location	CustomerName	Contact	ContactPhone	ExpectedAudience	PoliceDNR	PriorityName	Priority	Color	PreEstablishmentFromDate	PostEstablishmentToDate	EventFromDate	EventToDate	FromDate	ToDate
163	Manifestation mot Iranska regimen	Kungsportsplatsen -	IKF	Mehdi Amini	733,582,586	0	PRE/Hs-558-10/2013	LÃ¥g	3	#25E000	NULL	NULL	1/8/2014 16:30	3/31/2014 19:00	1/8/2014 0:00	3/31/2014 0:00

In this event, a protest was started 4:30pm the same day (i.e. 30 minute later than the speed drop observation).

6.1.5 Sensor specification data

Sensor technical and installation information data is selected from both STRESS1 and STRESS2 databases. The information is then fused into a single table of data that also includes geolocations, technical types of sensors, and lane ID if there are more than 1 sensors in a multilane road segment.

6.2 CLEAN DATA

The sensor data that results in speed measures over 250 km/h is considered error codes related and is removed from the data. Besides, the following data cleansing rules are apply:

- Records that cannot provide all required analysis data (during the analysis phase), i.e. null values presented.
- Records that have nonpositive speed measures.
- Records that have no traffic volume but still have speed measure information.

From the data quality assessment, we decide to only select the following 4 route segments for further processing: Kungsbackaleden (North and South), Tingstadstunneln (North and South).

The sensors representing these selected route segments are manually selected (from the geospatial view of sensor locations) as follows:

- Tingstadstunneln South: 911, 912, 7104, 7105, 7106, 7171, 7172, 7173, 7184, 7185, 7186
- Tingstadstunneln North: 7043, 7044, 7051, 7052, 7071, 7081, 7091, 7092, 7101, 7102, 7103, 7161, 7162, 7163, 7164, 7181, 7182, 7183
- Kungsbackaleden South: 1090, 1014, 1015, 1049, 1050, 1051, 7521, 7522, 7523, 7524, 7551, 7552, 7553, 7561, 7562, 7563, 7564, 7581, 7582, 7583, 7601, 7602, 7603, 7611, 7612, 7613, 7631, 7632, 7633, 7651, 7652, 7653, 7671, 7672, 7673, 7691, 7692, 7711, 7712
- Kungsbackaleden North: 1091, 1016, 1017, 1046, 1047, 1048, 7511, 7512, 7531, 7532, 7533, 7541, 7542, 7543, 7571, 7572, 7573, (7591), 7592, 7593, 7594, 7621, 7622, 7623, 7641, 7642, 7643, 7661, 7662, 7663, 7681, 7682, 7683, 7701, 7702

The following subgroups of sensors are also identified with data quality validation rules (numbers are sensor IDs):

- 911+912=7172+7173
- 7171+7172+7173=7184+7185+7186=7104+7105+7106
- 7071+7081+7091+7092=7101+7102+7103=7181+7182+7183=7161+7162+7163+7164
- 7091+7092=7052+7051=7044+7043
- 7521+7522+7523+7524=7551+7552+7553=7562+7563+7564
- 7561+7562+7563+7564=7581+7582+7583
- 7611+7612+7613=7631+7632+7633=7651+7652+7653=7671+7672+7673
- 7711+7712=7691+7692
- 7701+7702=7682+7683
- 7681+7682+7683=7661+7662+7663=7641+7642+7643=7621+7622+7623
- 7571+7572+7573=7591+7592+7593+7594=7541+7542+7543=7531+7532+7533

- 7511+7512 < 1046+1047+1048

6.3 CONSTRUCT DATA

To be able to link different datasets, we aggregate data by 5 minute intervals and reconcile the timescale (rounded to the same timescale) for all datasets.

Sensor data are first aggregated by the subgroup of the sensors, i.e. the sensors of the different lanes in the same route segment position, and then by the group of sensors (all subgroups of sensors representing the selected route segments). The aggregation is the average for speed and subgroup sum for the traffic volume and occupancy. For the analysis of severe congestion situation, aggregation using min/max is also used.

The unique key of the data rows is timeline (per interval of 5 minutes or 1 hour). The related information is:

- Measures: traffic flows at different road segments of interest.
- Events: weather, traffic situations.

This will allow non timeseries analysis to work. Trend/seasonal can be addressed by adding short-term memory of historical data (1 day or 1 week).

Only remove obvious outliers: volume <=0 or speed out of range <0,250>. Data rows that have null values in the required fields will be excluded during the modeling phase.

6.4 INTEGRATE DATA

Data are aggregated into only one single table indexed by equal intervals of 5 minutes. Aggregation rules are the average of the speed, sum of volume by lanes and average by segment, sum of occupancy.

Other independent variables (indexed by the same timestamp scale) are also included in the same integrated data table. The integrated variables include:

- Weather data: "AirTemp-C","RoadTemp-C","Daggp-C","Humidity-percent", "Wind", "Windmax", "WindDir", "NedbType", "Snow-mm", "Rain-mm", "Melted-mm", "TYta-Daggp-C", "RoadAirTemp-C"
- Road situation data: road situation (roadwork/accident) and accident information of the selected segments (kungälvsleden, kungsbackaleden, lundbyleden, and tingstadstunneln)
- Traffic flow measures: average speed, traffic volume (number of cars of type 0), occupancy (occupied time slot per measured time slot), traffic flow ratio (traffic flow in comparison with free flow traffic). Free flow traffics of selected route segments are calculated as the average of flows over period of 5:00-6:00am of all working days in 2014-2015.

Dimension information is also generated:

- Datetime: Timestamp information including both date and time of the measures
- Year: 2014 or 2015
- Day of year: 1-365
- Hour: 0-23
- Five minute block index: 1-12, indicating the 5 minute interval index in the hour.

- Working day (Monday to Friday and excluding Swedish holidays 2014-2015)

Figure 20 show the aggregated data.

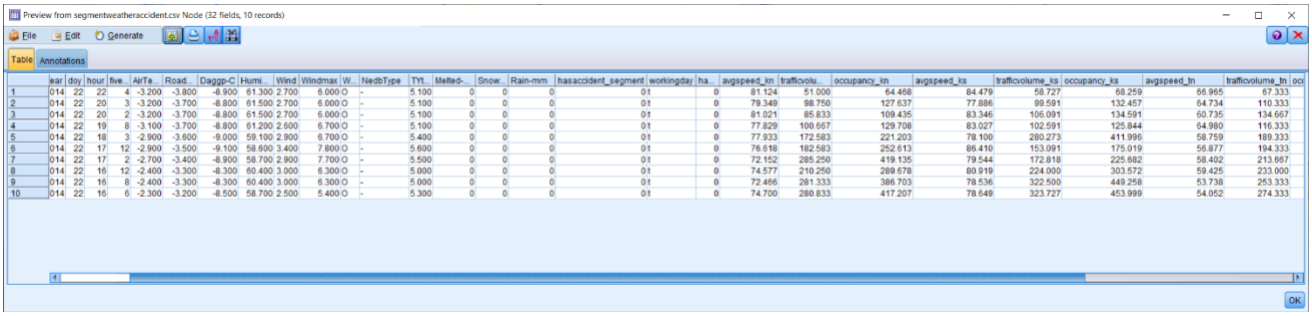

Figure 20: Aggregated traffic data with traffic/weather condition in 5 minute slot resolution

6.5 FORMAT DATA

Different data columns of the aggregated data table will then be formatted to specify type and role for the subsequent analyses. Since the data quality of the traffic flow measures is too low (only 11923/17520 hourly records have data), we do not have enough data to ensure the good results of timeseries based analyses. We therefore decide to consider the alternatives as fixed terms historical data analysis, where linear models are the most relevant methods to apply. Historical data of different historical data time slots are also gathered into the same data records. Figure 21 shows the aggregated data fields with related types, presented values and potential roles for the further analysis.

Traffic flow measures can be used for both purposes, e.g. as the input to predict the accident probabilities, or the output as predictions from traffic and weather conditions.

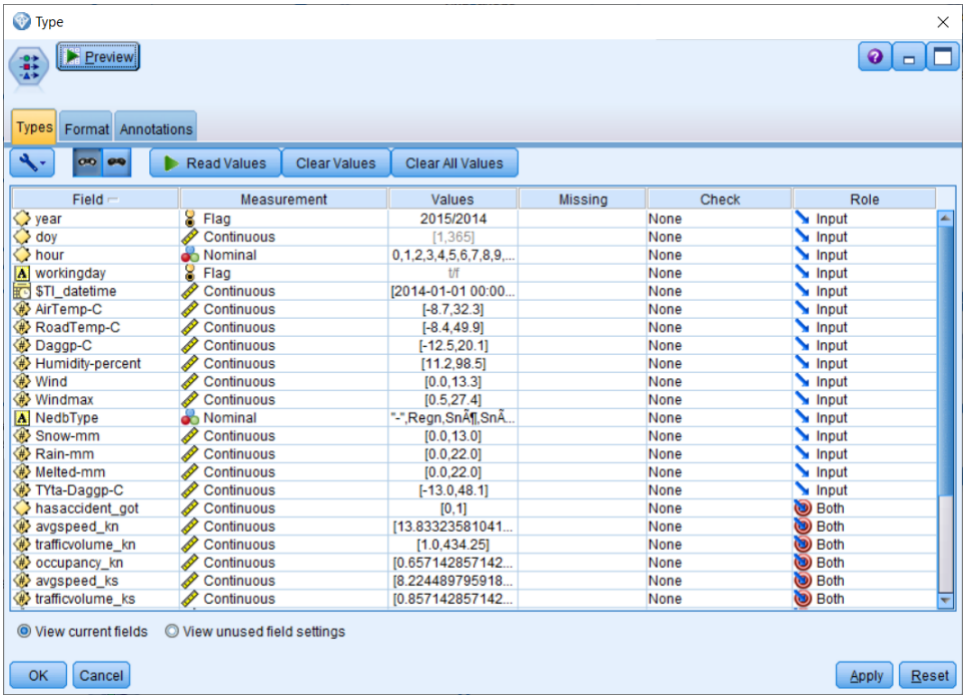

Figure 21: Formating data

7 MODELING

7.1 SELECT MODELING TECHNIQUE

The initial approach was to use timeseries analysis techniques where the dependent and independent variables are both described in form of time series data. However, because of the data quality of the collected STRESS1 datasets, timeseries analysis cannot promise to produce good results.

With the earlier observation from the data understanding that the traffic flow trend can actually be extracted from the most recent traffic flow (e.g. within couple of days), and that the flow patterns are related to working and non-working day, we then plan to analyse the potential correlation between traffic flow (speed), time of the day, weekday, weather, traffic situation and recent historical data of traffic flow.

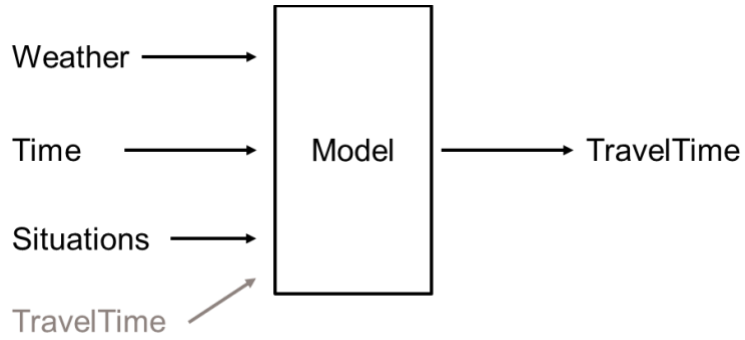

Figure 22: Data analysis models to be investigated

In this modeling phase, we thus decide to select the numeric class of models to predict numeric variables (travel time) as a function of the different set of input variables. The input variables are selected from the independent variables from weather, traffic situations, and also from the historical data of the output variables. Because of the nonlinearity relationship between traffic flow and the time of the day, the hour index can be used to split linear models applicable for different time of the day.

Numeric class of models supported by SPSS and Watson modeler consist of:

- Tree based
  - Random trees
  - Tree-AS
  - CHAID
  - C&R Tree
- Linear regression
  - Regression
  - Generalized Linear

    - Linear-AS
- SVM
    - SVM
    - LSVM
- Neural network
- KNN

The following sections will describe several experiments with different model techniques in the list where the results are considered best among the available models. Ranking evaluation of models are based on either correlation, number of required input variables of relative errors.

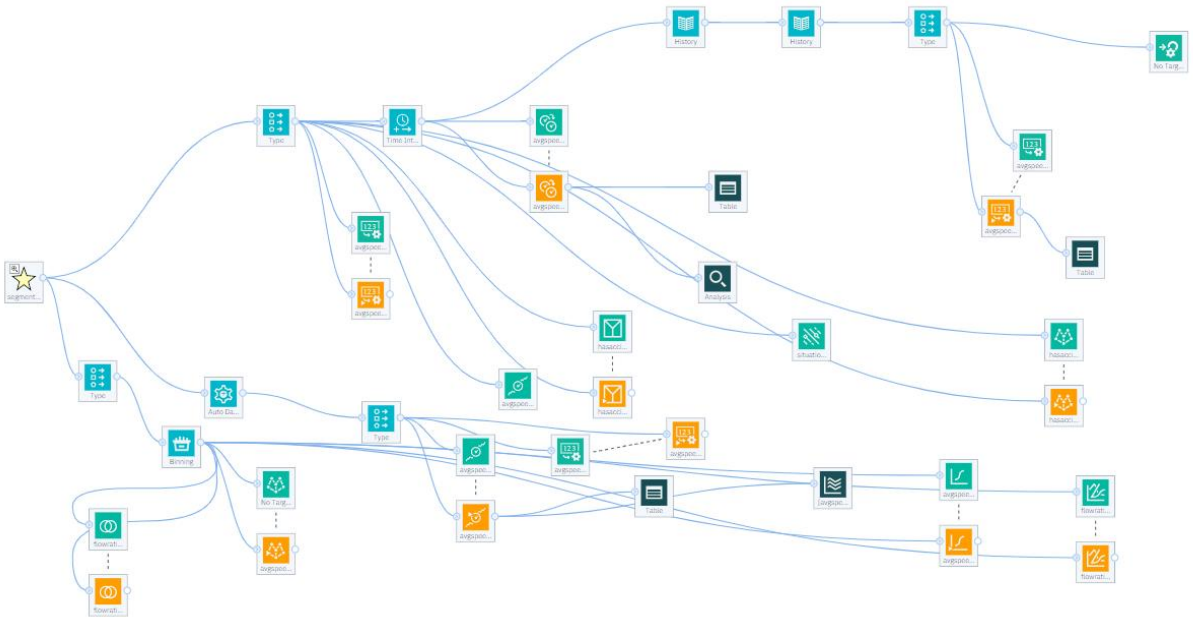

Figure 23: Watson models used in the experiments

7.1.1 C&R tree analysis

The Classification and Regression (C&R) Tree is a tree-based classification and prediction method. This method uses recursive partitioning to split the training records into segments with similar output field values. The C&R Tree starts by examining the input fields to find the best split, measured by the reduction in an impurity index that results from the split. The split defines two subgroups, each of which is subsequently split into two more subgroups, and so on, until one of the stopping criteria is triggered. All splits are binary, meaning that there are only two subgroups.

In this experiment, we assume that the traffic flow is dependent on type of days (working or nonworking), hour of day, and severe weather conditions (snow, rain). The C&R tree model is implemented in SPSS modeler (Watson modeler) and resulted in two separate decision trees as illustrated in Figure 24 and Figure 25 for working days and nonworking days respectively.

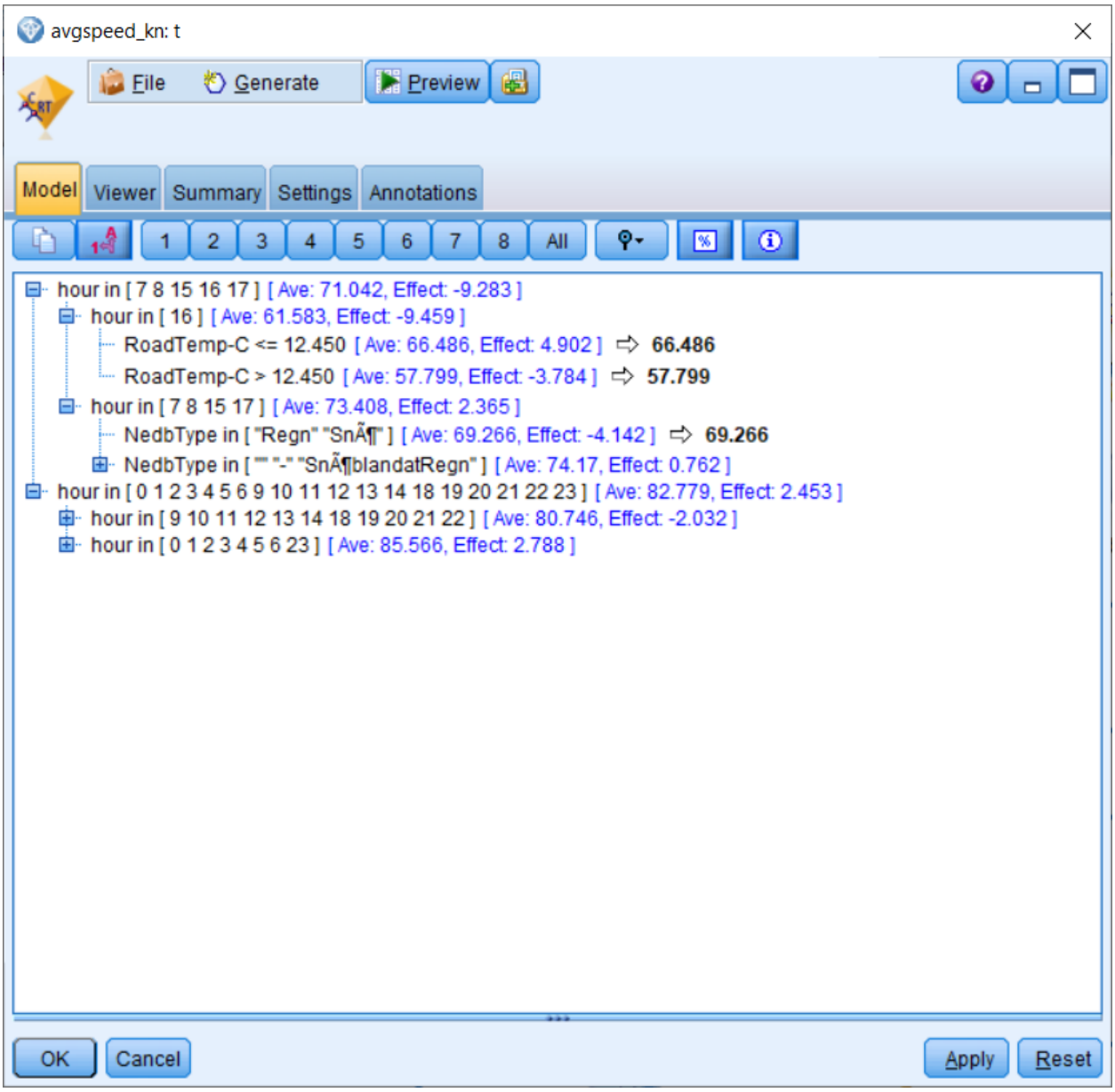

Figure 24: C&R tree analysis results of Kingstadstunneln North for working days

The analysis shows that the detected peak hours are 7:00-8:00 and 15:00-17:00 (working day) and during peak hours, severe weather conditions introduce a noticeable decrease in average speed of the selected segment.

In nonworking day (weekend or holiday), there are no rush hours but the segments are mainly morning 1:00-9:00 and the rest of day.

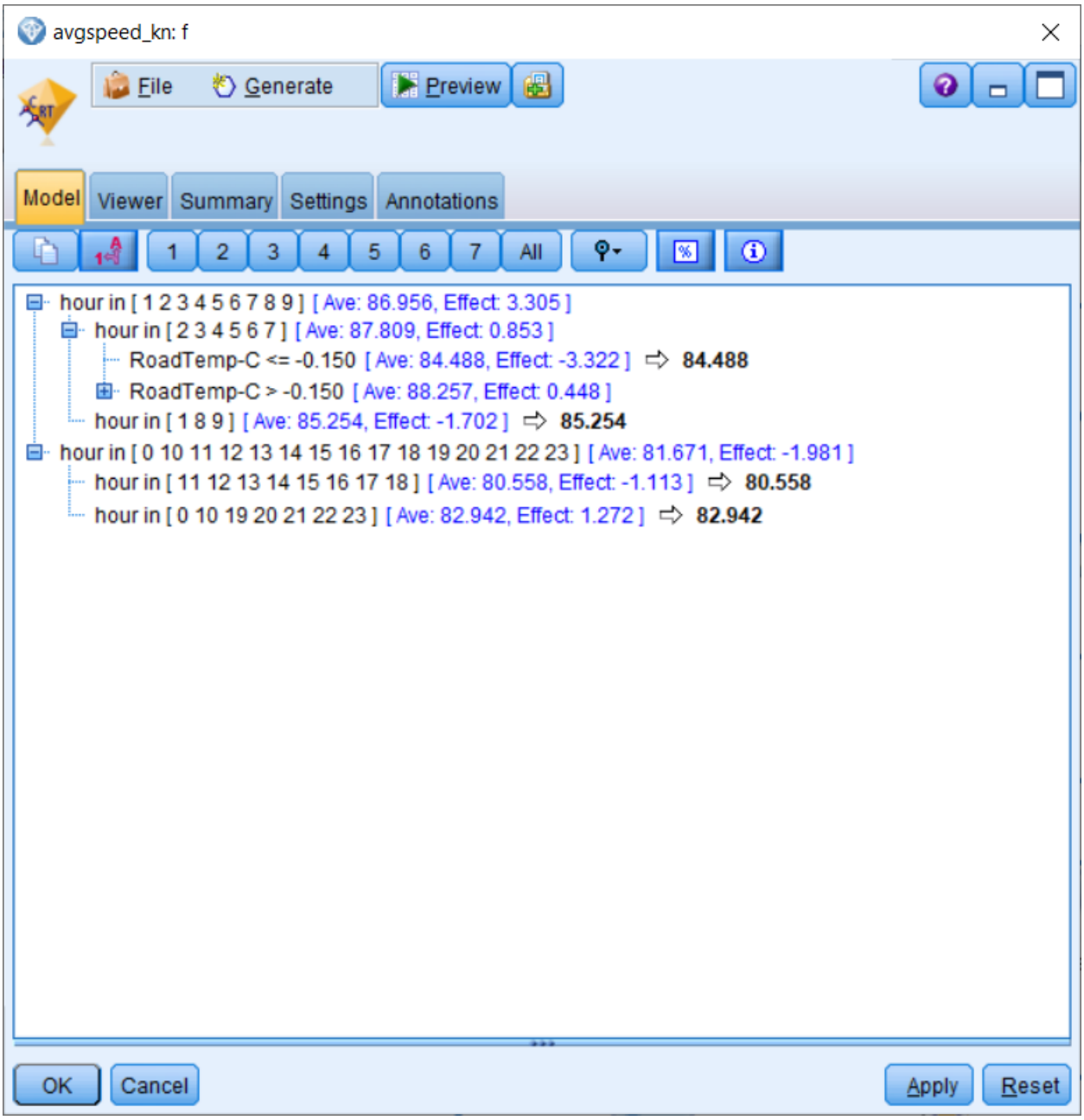

Figure 25: C&R tree analysis results of Tingstadstunneln North for nonworking days

7.1.2 Accident analysis

In this series of experiments, we want to explore the correlation of accident instances as a result of weather, time and traffic flow measures. The experiments are performed with classification class of models (Tree-AS, CHAID, LSVM, etc…). Results of Linear SVM on the predictor importance are shown in **Error! Reference source not found.**. Interpretation of the result is that the accidents occurred in the Kungsbackaleden route segment are not very much correlated to weather parameters except severe weather condition, but instead depended more on the traffic load at Kungsbackaleden and time of the day. However, this hypothesis should be further refined and validated when we can have more collected data to provide highly accurate conclusions.

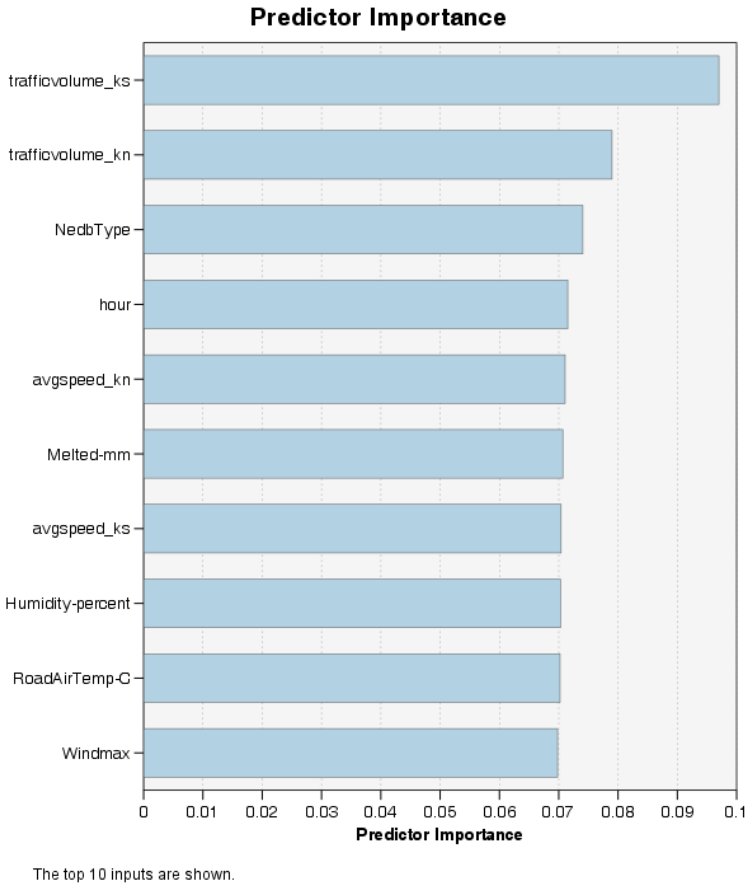

Figure 26: Linear SVM predictor importance on prediction of accidents

7.1.3 Speed and congestion forecast

Numerical class of models are used to find the correlation of average speed (over 5 minutes or 1-hour sampling intervals) with other input parameters: weather, road situation, time period (hour of day or working/holiday) as well as the historical speed measured at the selected routesegment. Further experiments can also be performed to find the correlation of speed of a routesegment with the historical speed of other route segment(s). This is based on assumption that traffic patterns of commuters are dominant and predictable.

Different numeric models are used for this series of experiments (see Figure 27). All these models are executed against the same set of input and output variables. Comparison of the model results and the clarity of the models leads to the selection of regression model. Therefore regression models will be selected for further analyses to derive the forecasting functions for the deployment.

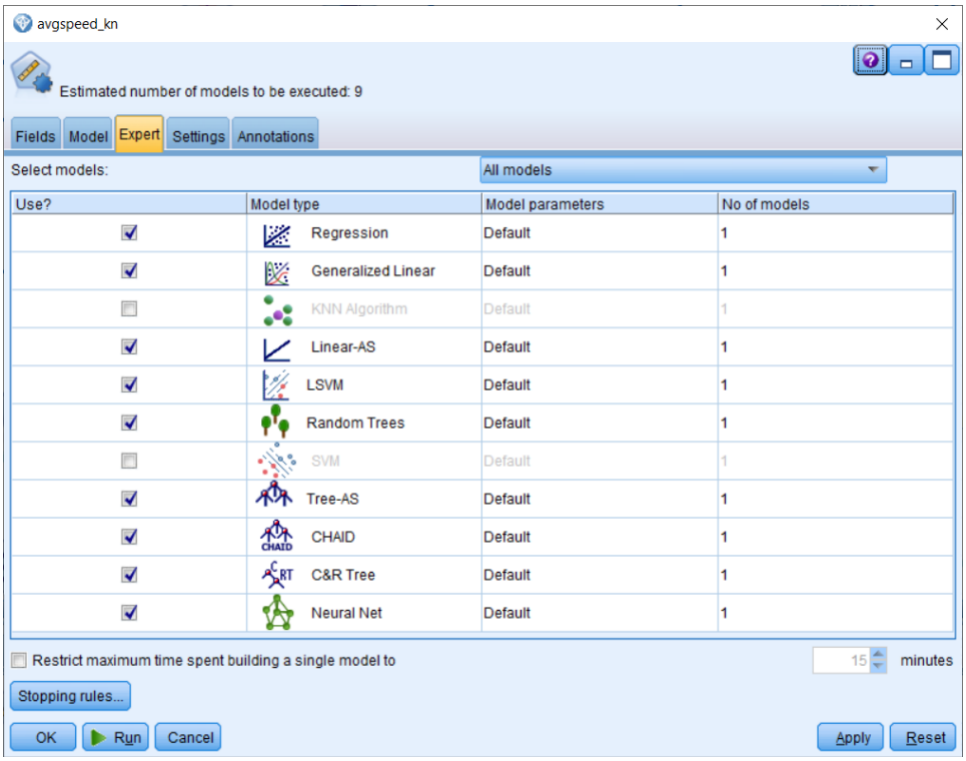

Figure 27: Numeric models to predict average speed by using different methods

7.2 BUILD MODEL

Analytical models are built using the SPSS modeler (and Watson data modeler). The desktop version SPSS and the cloud version Watson studio are using the same model file which can upload and/or download depending on the analysis need. Figure 28 illustrates the models designed in SPSS desktop and uploaded to Watson studio in IBM Cloud. The model starts with input node that will access the input aggregated data (Section Integrate data6.4) either as a single CSV file or a table in PostgresSQL. Data format (as described in Section 6.5) was defined in Type node in the sequence. The same formatted data as output from this node will be used for all different models (Figure 28 shows two model classes using the same input formatted data).

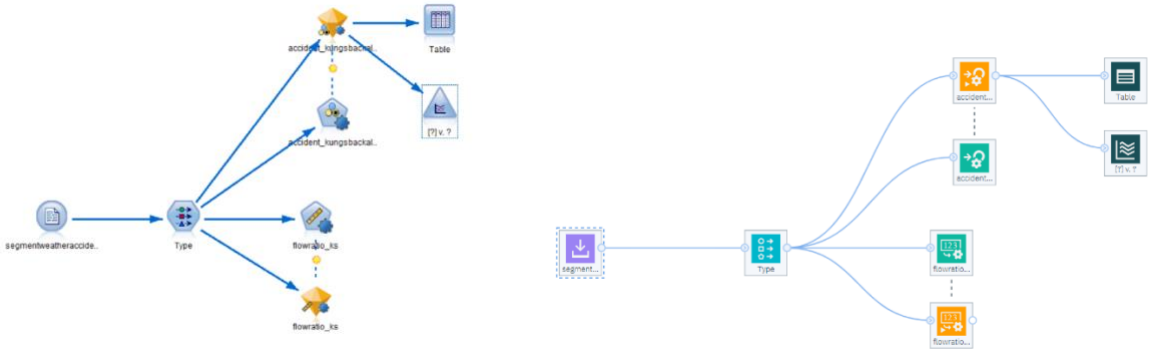

Figure 28: Different model classes in the same SPSS (left) and Watson studio (right) model stream

Historical data of traffic flow was created on the fly via first using time intervals to create the data rows evenly distributed over timescale with intervals of 5-minute or 1-hour, and then using the history node where a number of lag data will be created as addition attributes into the aggregated input data.

After adding these historical fields into the input aggregated data, models are created similar to the previous experiment, except that now start from the history node output. Figure 29 shows different models using different time intervals and different history settings.

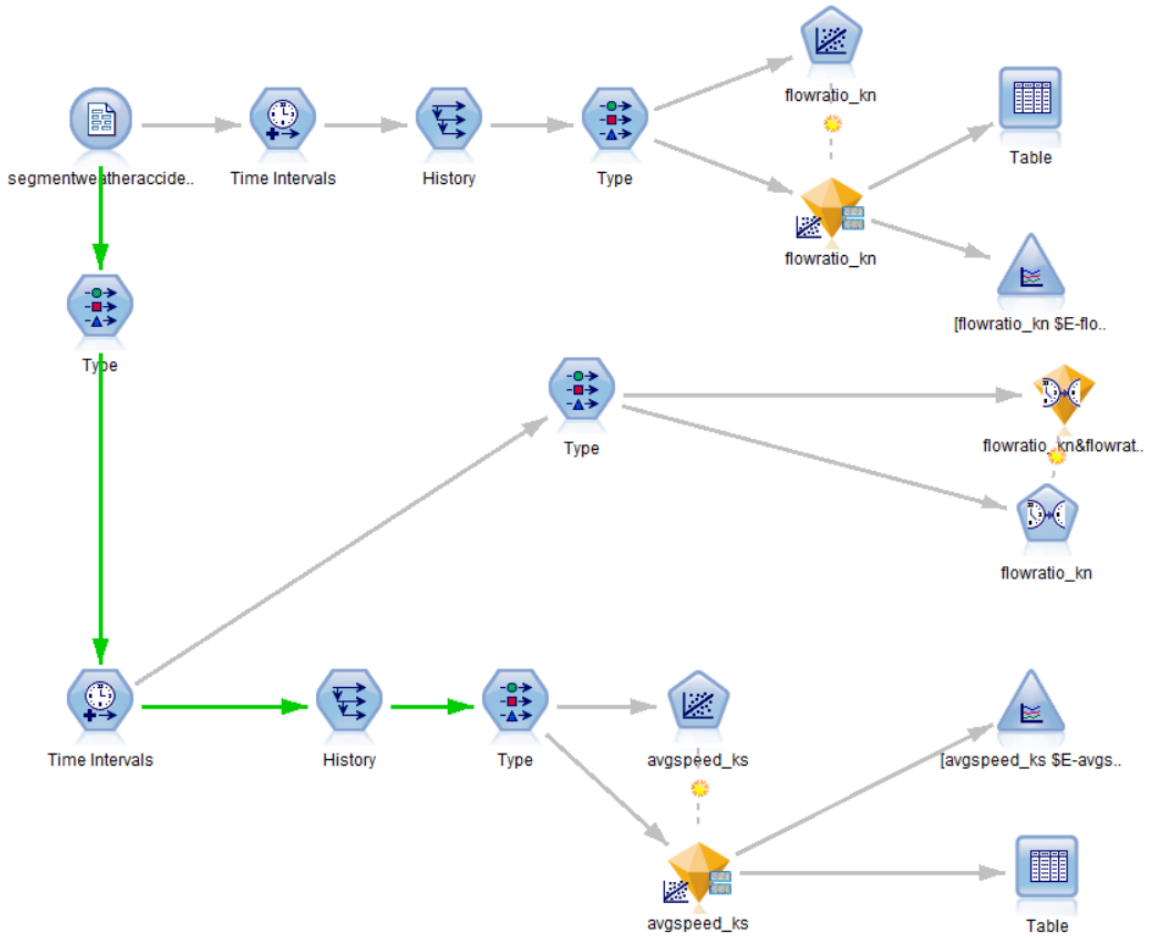

Figure 29: Models with history node and time interval

After building the model, we run the selected stream in the model (green path in Figure 29) that consists of linear regression model with 5-minute interval historical data. The model is set to provide different models for working and nonworking day forecasting of average speed at Kungsbackaleden South. Figure 30 shows the result of model run, where two models are found with the correlation of 98.8% and 98.6% for nonworking and working days respectively.

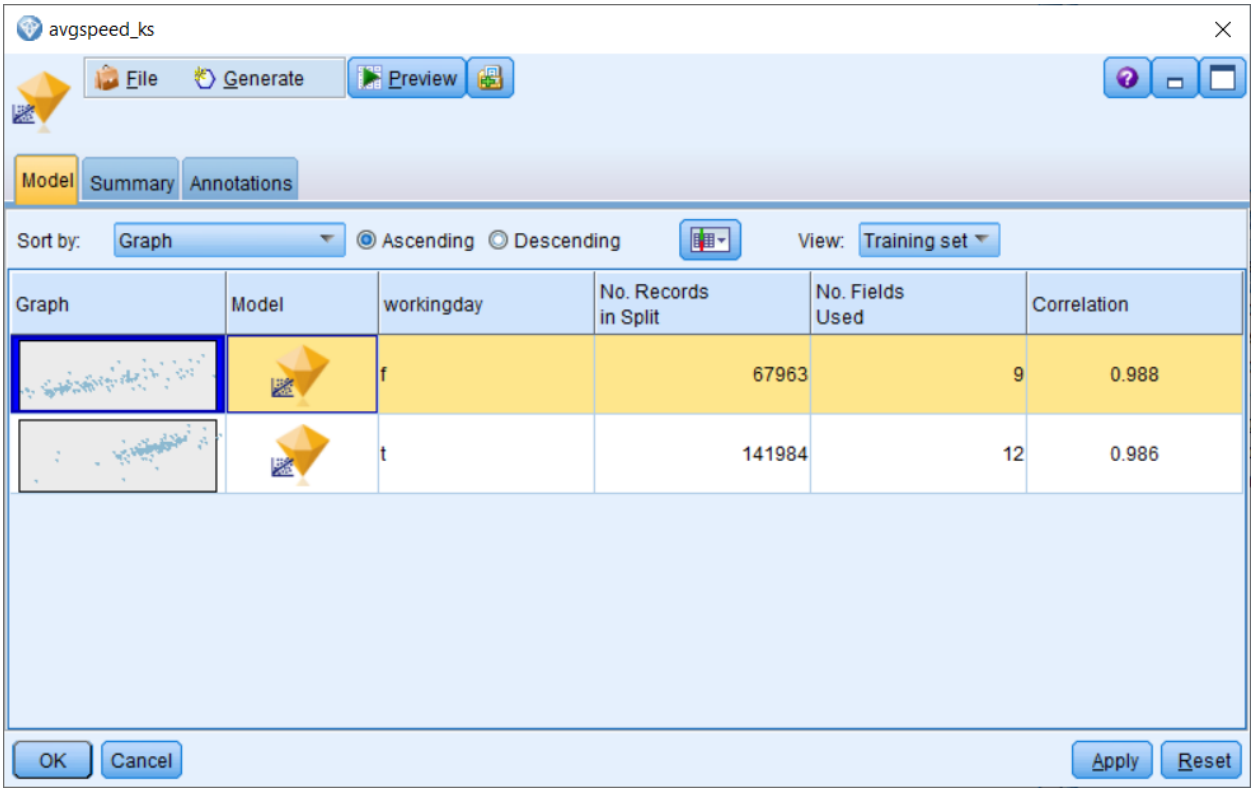

Figure 30: Summary result of linear regression model

Details of found forecasting models are provided in Figure 31and Figure 32 as a linear function of input variables (weather parameters and the last 5 lags of historical 5-minute interval data of road situation and average speeds).

Similar experiment had also been performed with 24 lags of historical 1-hour interval data (i.e. one day forecast). Figure 33 and Figure 34 show the found forecasting models for working and nonworking days. The correlations now are slightly lower than previous experiment (90.5% and 80.9% for working and nonworking day respectively) and the time resolution is 1 hour (same forecasted value for 1-hour interval).

These two experiments can be setup with very little cost, the only difference is in the settings in history node (24 lags of 1-hour) and the selection of input/output variables. Similar experiments can be established with the same approach for different hypothesis and management policy. For the sake of simplicity while still not losing the generality of the approach, we only select the 5-minute interval experiment for next steps in the lifecycle.

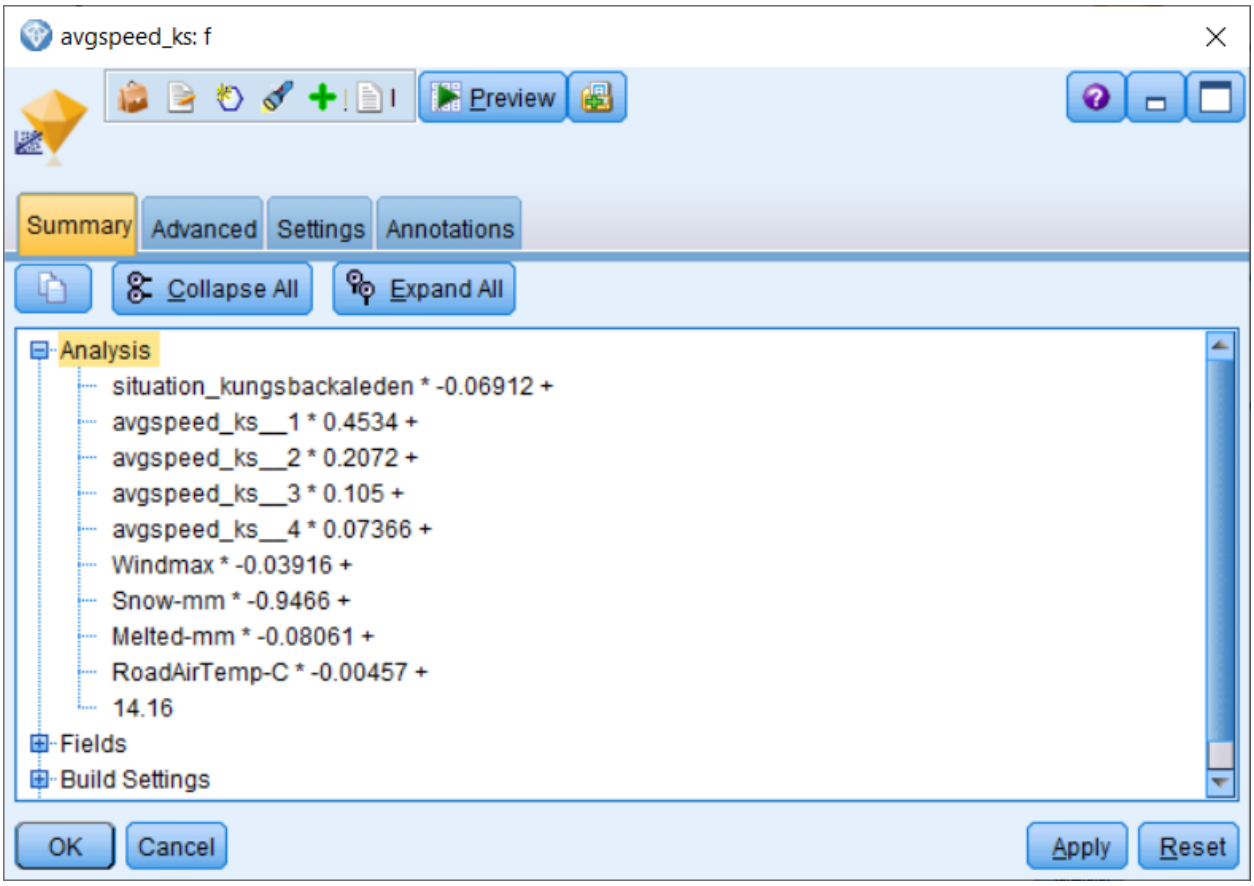

Figure 31: Forecasting model for average speed Kungsbackaleden South (nonworking day)

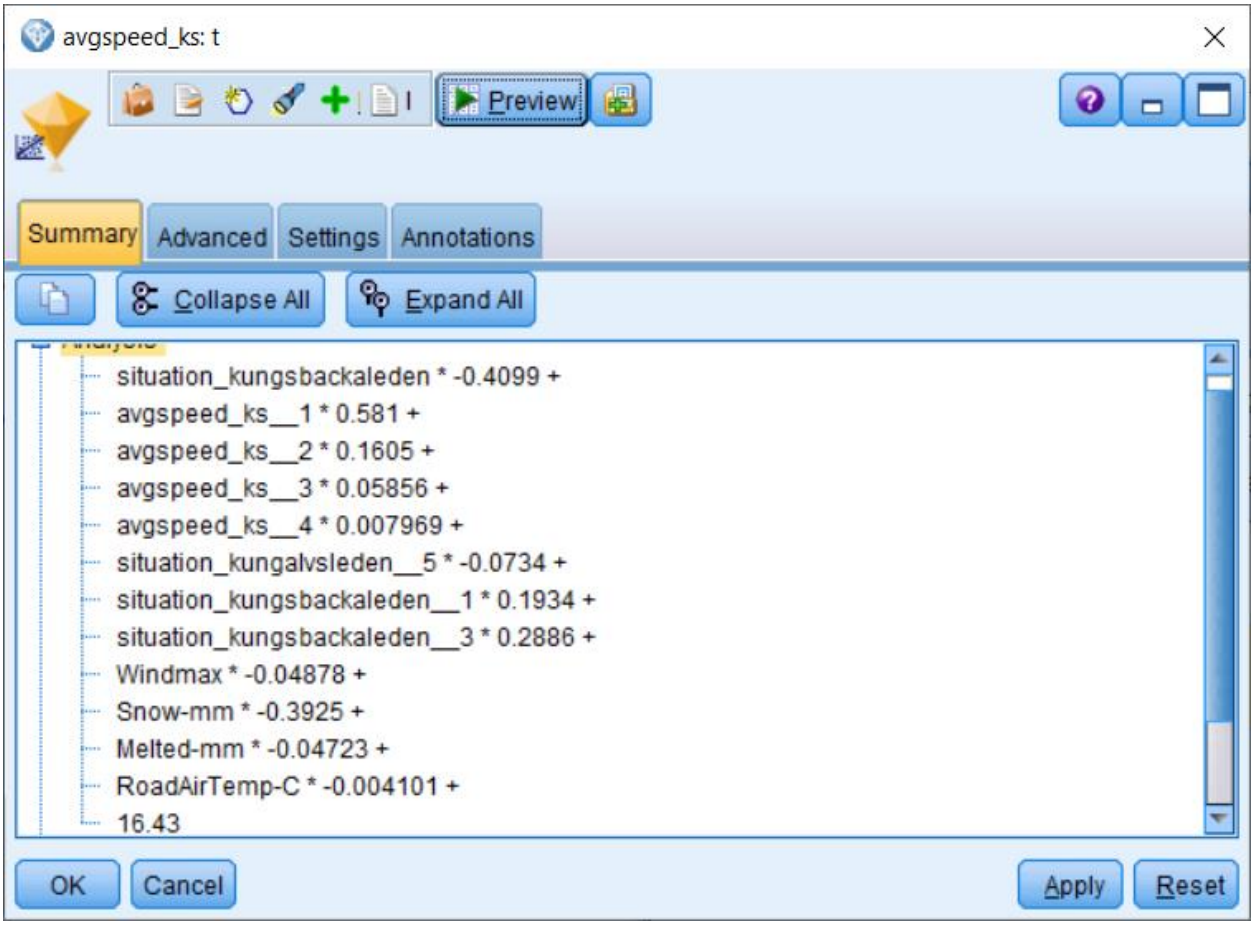

Figure 32: Forecasting model for average speed Kungsbackaleden South (working day)

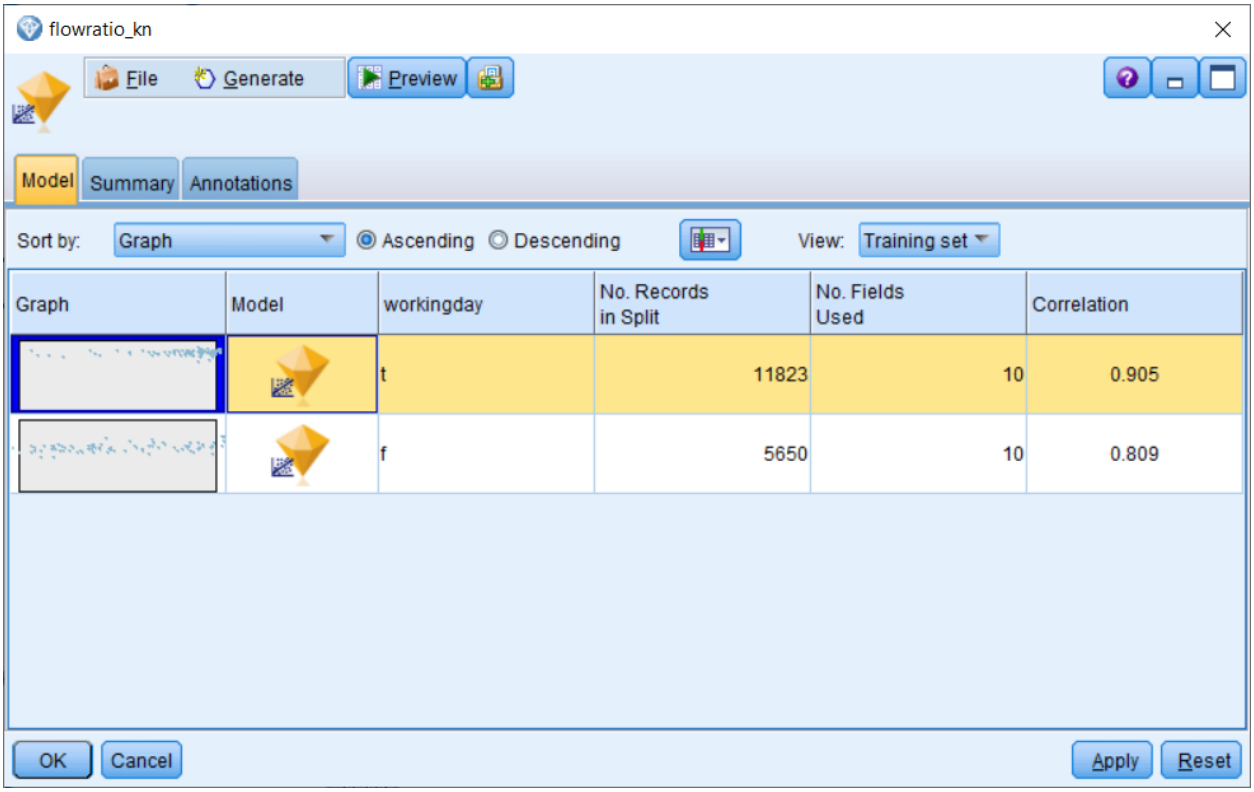

Figure 33: Forecasting traffic flow Kungsbackaleden North from 24 hour historical data

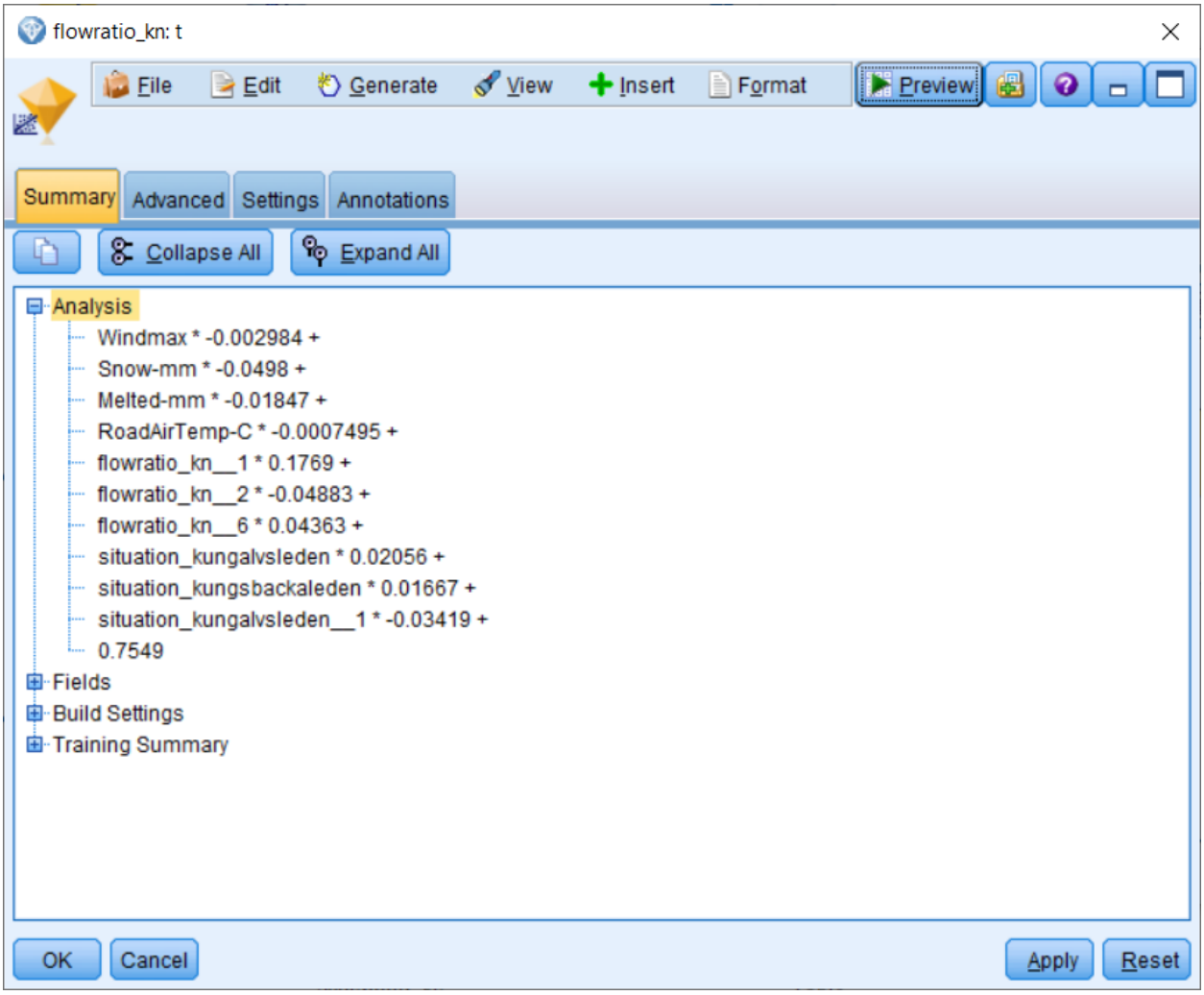

Figure 34: Forecasting model of traffic flow Kungsbackaleden North based on 24hours data

7.3 ASSESS MODEL

Models from each class will then be compared by different criteria. Figure 35 show the results of the same experiment, i.e. the prediction of the average speed in Kungsbackaleden North, but using different models of choice. The assessment is based on correlation, number of input fields used and the relative error rates.

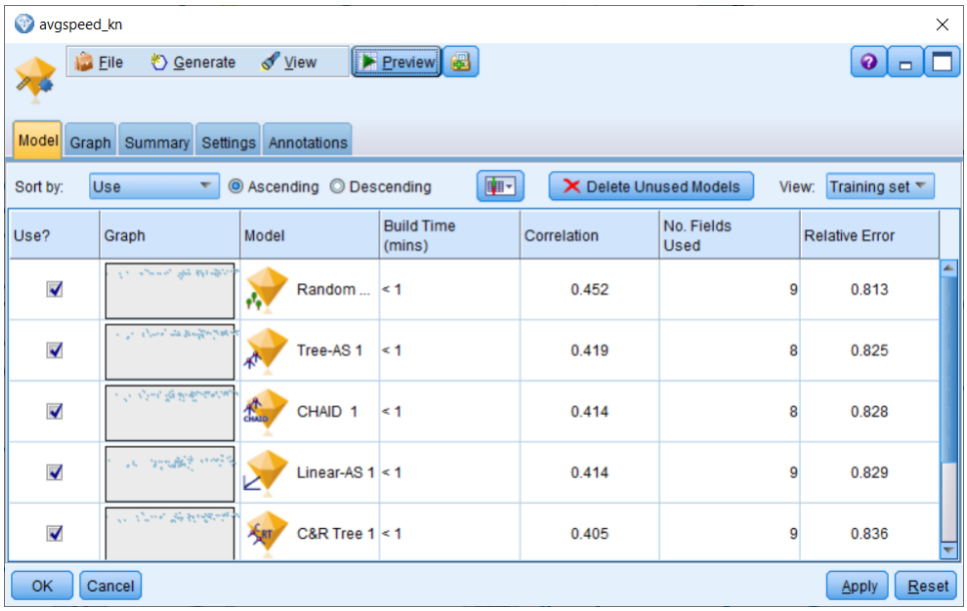

Figure 35: Comparison of different selected method in estimation accuracy

We assess the rules found by different models in the list, the achieved accuracy (**Error! Reference source not found.**), and the simplicity for the prototype deployment (to be described later in the report) with selected technology. The assessment resulted in the selection of the forecasting models in Figure 31 and Figure 32 for further analysis and deployment phase.

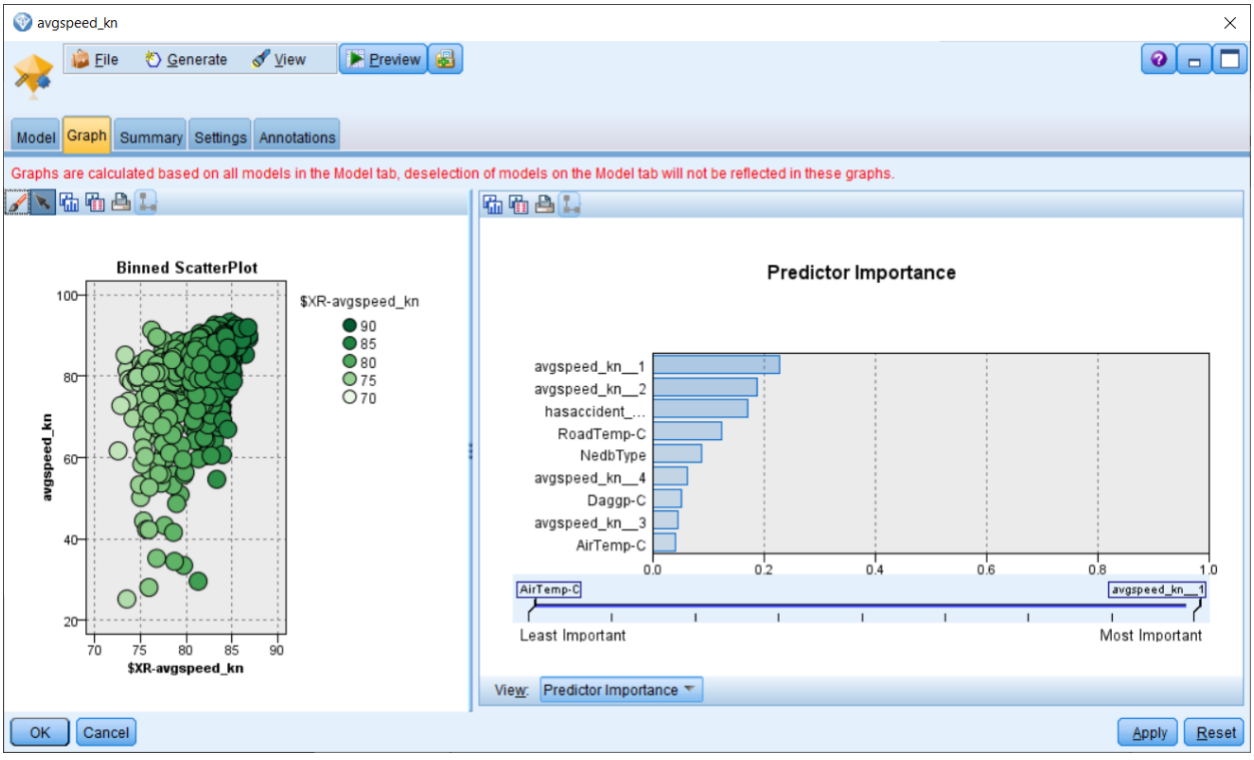

Figure 36: Predictor importance and estimation vs real data plot

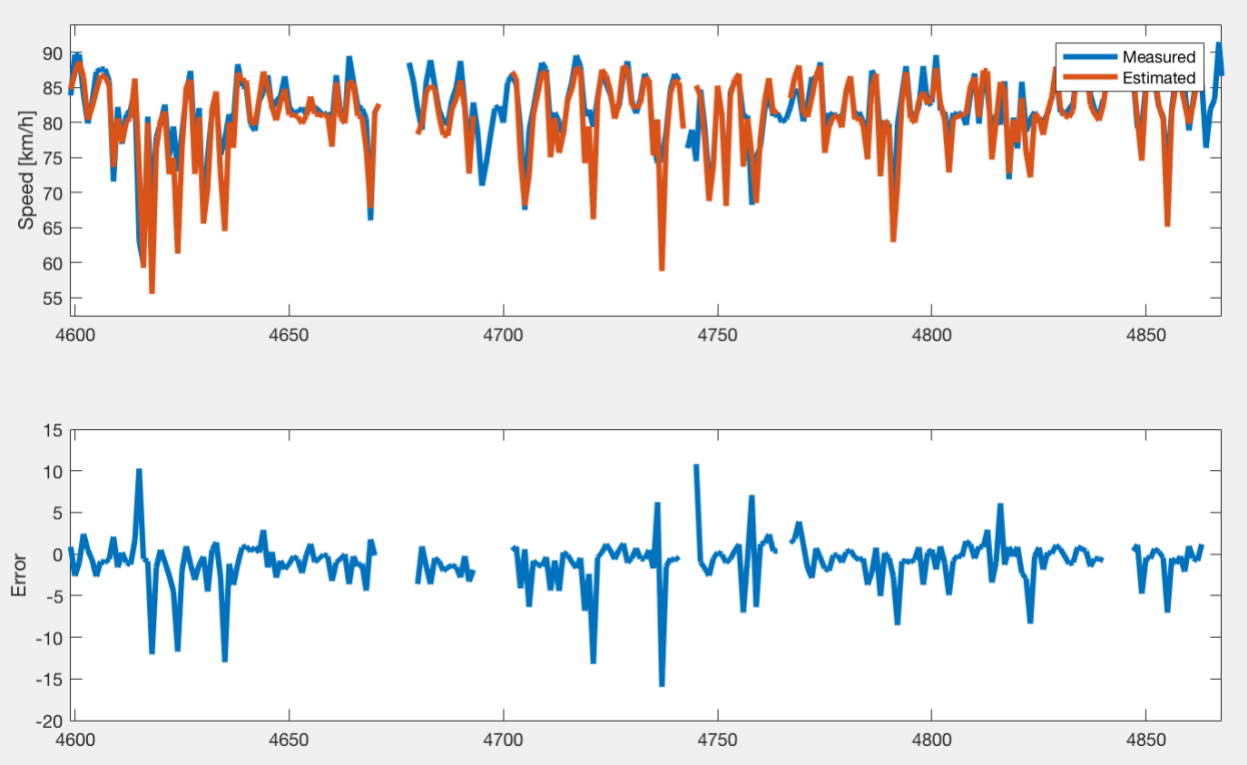

Figure 37: Linear regression with 24 hours historical speed data

Error! Reference source not found. and Figure *37* illustrate the assessments based on different perspectives of forecasting quality measures (relative errors of estimated vs actual, predictor importance, and correlation between estimated and actual data).

8 EVALUATION OF THE RESULTS

Throughout different lifecycles of data mining, the following findings are found to be valuable:

- The data quality has potential to be improved for data analysis purposes. This fact restricts the flexibility of analysis models to be applied. For example, time series analysis cannot be used. For data driven traffic management, a proper data management framework should be in place at STA. Datasets existing today at STA are mainly proprietary assets of different legacy subsystems used for their own business purposes. It is crucial to have a separate data management process to catalogue the existing data, identify data quality needs and improvement actions from different analysis initiatives, etc. that can serve data driven approaches.
- Data driven approaches can also derive similar traffic management decisions (e.g. peak hours, weather impacts, etc.) as human being from experiences. This will help STA with a systematic approach for decision making.
- Different techniques have been used to linearize the problem, and thus being able to analyze with simpler linear models with more clarity and easier to implement.
- The requirements of the data quality are dependent on analysis objectives. Therefore, the data quality improvement process is recommended to be a daily process if STA wants to deliver more values from data insights.

- Data driven mindset: The data driven mindset should be adopted in different projects. This will avoid waste of efforts collecting bad quality data that may not be useful for later analysis.
- The more datasets that are collected and integrated, the more hypotheses can be validated.
- Understandings of the business and the data can be constantly improved in correlation to each other steps.
- The possibility developing models locally based on a smaller data set and, then evaluating the full data sets in the cloud is a benefit for the development process.

9 DEPLOYMENT

9.1 PROTOTYPE WEB APPLICVATION

A prototype web application is available at http://datapolicy-map.mybluemix.net/ presenting the predicted traffic speed at different road segments based on historical traffic flow, weather, an traffic situations. The purpose of this development is to illustrate how the methodology presented in this report can be used in the future when a more complete data set is used in the evaluation. The application is built using a Python Flask backend deployed on IBM Cloud and connected to a PostgreSQL server, also deployed on IBM Cloud.

A model for forecasting the free flow along Kungsbackaleden is provided by the resulted model(s) as described in Section 7 and incorporated into an API (see below). In order to visualize the forecast, the web application uses an open source map library called Leaflet (https://leafletjs.com). Four road segments are visible (Tingstadstunneln, Kungsbackaleden, Kungsälvsleden, and Lundbyleden), but only one (Kungsbackaleden) is forecasted. A call to the API yields the forecasted free-flow ratio, i.e. the forecasted speed divided by the free-flow speed, along Kungsbackaleden. A free-flow ratio of 1 indicates that there is no slow-down on the road.

Following the same rules as Trafiken.nu, the road segment is colored:

- **Dark red:** $0.15 \geq$ forecast ratio
- **Red:** $0.30 \geq$ forecast ratio > 0.15
- **Orange:** $0.70 \geq$ forecast ratio > 0.30
- **Green:** forecast ratio > 0.70

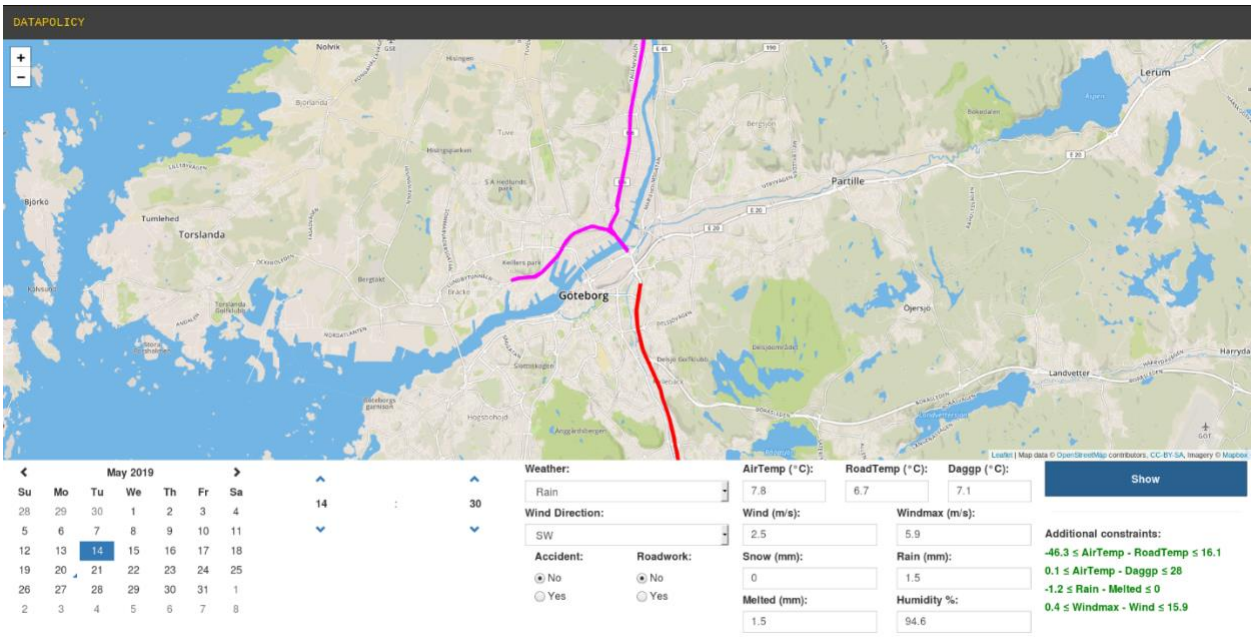

Figure 9: Example forecast yielding a red road segment, indicating heavy traffic.

Based on historical data, constraints are placed on each value a user can input. If these constraints are not met, the violating value/constraint will turn red and no forecast can be made until corrected.

9.2 PROTOTYPE FORECAST API

The model provided for forecasting the traffic freeflow ratio along Kungsbackaleden has been implemented as an API.

9.2.1 Usage

The API is called through

http://datapolicy-map.mybluemix.net/api/flow/{day_of_year}/{hour_of_day}/{five_minute_segment_of_hour}?weather={weather_nr}&winddir={wind_direction_nr}&accident={accident}&roadwork={roadwork}&roadtemp={road_temperature}&airtemp={air_temperature}&daggp={dew_point}&snow={snowfall}&rain={rainfall}&melted={melted}&windmax={max_windage}&wind={wind_speed}&humidity={humidity_percentage}&weekend={weekend}

where each curly bracket enclosure should be replaced with an appropriate value for the corresponding parameter:

Parameter	Value
{day_of_year}	The day of year to forecast. January 1 is day of year 1.
{hour_of_day}	The hour of day to forecast, starting with 0.

{five_minute_segment}	The five minute segment to predict, starting with 0 (on the hour).
{weather_nr}	0=none, 1=rain, 2=snow, 3=mixed.
{wind_direction_nr}	0=none, 1=north, 2=north-east, 3=north-west, 4=east, 5=south, 6=south-east, 7=south-west, 8=west.
{accident}	1 if accident, 0 if no accident.
{roadwork}	1 if roadwork, 0 if no roadwork.
{road_temperature}	The temperature of the road in ℃.
{air_temperature}	The temperature of the air in °C.
{dew_point}	The temperature at which the air must be in order to be saturated with water vapor, in ℃.
{snowfall}	How much snow has fallen, in millimeters.
{rainfall}	How much rain has fallen, in millimeters.
{melted}	
{max_windage}	The maximum speed of the wind in meters per second.
{wind_speed}	Speed of the wind in meters per second.
{humidity_percentage}	Humidity percentage.
{weekend}	1 if it is a Saturday or Sunday, 0 otherwise.

9.2.2 Response

When calling the API, the forecasted free flow ratio along Kungsbackaleden is received in json format.

Example API call:

http://datapolicy-map.mybluemix.net/api/flow/140/11/12?weather=1&winddir=7&accident=0&roadwork=0&roadtemp=6.7&airtemp=7.8&daggp=7.1&snow=0&rain=1.5&melted=1.5&windmax=5.9&wind=2.5&humidity=94.6&weekend=0,

results in the following response:

{

"kungavls_north": -1,

"kungavls_south": -1,

"kungsbacka_north": 0.9351232246845382,

"kungsbacka_south": 0.9351232246845382,

"lundby_east": -1,

"lundby_west": -1,

"tingstad_north": -1,

"tingstad_south": -1

}

As we can see, the forecast along Kungsbackaleden for these particular parameter settings is that we have 93.5% of the free flow. The key-value pairs whose value is –1 are unimplemented forecasts and should be ignored.

10 CONCLUSIONS

The aim of the project was to take the first steps to describe how traffic authorities can experiment with data-driven policy development. The process finding a suitable question to be analyzed covered many different areas, and gave lots of insights in the kind of issues related to the policy data driven field. One such aspect is the importance as a data analysist of being involved in the data collection phase, and not only the data analysis phase. This covers what sensors that are used, follow up so data is continuously stored, but also what information that is stored.

Based on the data available it is difficult to predict the traffic situation without using the current traffic situation. This implies that the prediction time is hours rather than days. If the prediction horizon is wanted to be days instead, more data or a data-driven approach combined with a model-based approach may be beneficial.

A generic palette of tools and analysis is developed. These different methods can be used for increased data-sets or different applications. To illustrate how the results can be used in a more user friendly way, a prototype map application presenting the predicted traffic flow is developed.

11 FUTURE WORK

Based on the process and the results, several ideas for future work is formed and described in this section.

- One idea is to find times where there are significant differences between the estimated traffic speed and the measured, and investigate whether there are some explanations for this difference in the data sets available. This will hopefully increase the understanding of what is affecting the traffic situation, even if these situations does not occur on a frequently basis.
- Use an increased data set. Examples of this could be to use GPS-signals from the traffic or use a longer time period of STRESS data.
- Combine the data driven analysis with a model based approach and investigate how much the prediction accuracy is increased.
- Use the same framework for other applications, e.g. the railway system.
- Extend the data sets to also include information about how the traffic situation is affected by building constructions in a city.

12 REFERENCES

C. Schrader, C., Kornhauser, A., & M. Friese, L. (2004, 1). Using historic information in forecasting travel times.

Domenichini, L., Salerno, G., Fanfani, F., Bacchi, M., Giaccherini, A., Costalli, L., & Baroncelli, C. (2012, 10). Travel Time in Case of Accident Prediction Model. *Procedia - Social and Behavioral Sciences, 53*, 1078-1087. doi:10.1016/j.sbspro.2012.09.957

Hojati, A. T., Ferreira, L., & Charles, P. (n.d.). Assessing the major causes of travel time reliability on urban freeways. 10.

IBM. (2011). IBM SPSS Modeler CRISP-DM Guide.

Koesdwiady, A., Soua, R., & Karray, F. (2016, 12). Improving Traffic Flow Prediction With Weather Information in Connected Cars: A Deep Learning Approach. *IEEE Transactions on Vehicular Technology, 65*, 9508-9517. doi:10.1109/TVT.2016.2585575

Lint, J. W., Willem, J., Lint, C. V., Voorzitter, R. M., Dr, P., Ir P. H., L., . . . Katholieke, H. (2004). *Reliable Travel Time Prediction for Freeways.* Tech. rep.

Parker, A., Simari, G. I., Sliva, A., & Subrahmanian, V. S. (2014, 1). *Data-driven Generation of Policies* (2014 edition ed.). New York: Springer.

QGIS Development Team. (2019). QGIS Geographic Information System. Open Source Geospatial Foundation Project. Retrieved from http://qgis.osgeo.org

Tu, H., Lint, H. W., & Zuylen, H. J. (2007). Impact of Adverse Weather on Travel Time Variability of Freeway Corridors. Retrieved from https://trid.trb.org/view/801786

Wirth, R. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining,* (pp. 29-39).

Yildirimoglu, M., & Geroliminis, N. (2013, 7). Experienced travel time prediction for congested freeways. *Transportation Research Part B: Methodological, 53*, 45-63. doi:10.1016/j.trb.2013.03.006

13 APPENDIX 1: INTERVIEWS WITHIN STA

As a part of the work towards finding a suitable area for demonstration and prototyping, a number of interviews was conducted with representatives from STA. In what follows, the results of these interviews are presented.

13.1 POTENTIAL AREAS

In the following, some potential application of data-driven policy work, found in the interviews are presented:

- **Aggressive road traffic:** some drivers move overly aggressive in traffic and cause a large portion of traffic accidents. By using advanced image analysis, it might be possible to identify such behavior. In addition, it could be possible to monitor whether actions taken towards minimizing such aggressive driving have the desired effects.
- **Recommended departure times:** given the vast historic data on travel times, it should be possible to provide necessary departure times the day before (given weather forecasts, events and other known or probable facts).
- **Snow clearance:** Collect data from OEMs about current road conditions – in order to both follow-up whether snow plowing contractors are performing sufficiently and to be used as a source of traffic information.
- **Train Traffic Information Accuracy:** Currently, departure and arrival times do not meet customer requirements. Several interesting viewpoints on the subject was collected and accounted for here
    - One major challenge in arriving at a correct travel information is the long chain of actors necessary to produce correct arrival and departure estimates. STA only produces 13-18% of the necessary information, operators, contractors and other supply the rest. The main challenge lies in gathering the data, a lot of communication is done over the phone and is never recorded.
    - Another challenge has been the uneven pace in the projects. Initiatives are started, terminated, and when rejuvenated technology har surpassed, past results and pre-studies must be done all over. The biggest project is currently NTL (new traffic management system).
    - Another barrier to more data-driven practices concerns getting high-quality real-time data. Historical data typically loses relevance over time as it is recalculated and changed.
    - One identified need is to work with more proactive visualization about ongoing and future train operations. STA has strong capabilities when it comes to visualizing historic data – but is lacking when it comes to real-time and future predictions.
    - Another area concerns analysis of both numeric data and deviation reports (text). In combination these two data sources could serve as a basis to provide answers to delays in the system. With new technology this should be possible.
    - Finally, one area concerns predicting the consequences of common system perturbations. It should be possible to predict the consequences of e.g. switch breakdowns.
- **Autonomous vehicles' impact on the transport system:** There is an imminent risk that autonomous vehicles can come to challenge public transport as we know it today. The

reason public transport is financially sustainable today is that many people choose public transport although they could afford traveling by other means. In the case that autonomous vehicles were to have widespread market penetration, there is a substantial risk that many travelers choose this and other sharing services over public transport. This would risk to hollow the basis public transport and those who lack the opportunity to choose other means of transport will thereby be even more segregated. It would be very useful if it was possible through data-driven analysis to simulate the effects on public transport in case of higher usage of autonomous vehicles and other sharing services.

- **Acceptance of geofencing:** It is very likely that cities and road authorities will increase the use of geofencing – i.e. how authorities through digital tools and infrastructure can control and manage traffic in certain area remotely. This can be due to ensuring low speed near schools or eliminate the risk for terrorist attacks at central and densely visited areas. Since geofencing marks a new, more coercive style of traffic management, using data-driven simulation to explore user acceptance of geofencing would be very useful.

13.2 Information Security Considerations

- Information security is a high priority area within STA
- All information objects within STA must be classified. There are several thousands of such objects within STA.
- This has not been done to a sufficient extent – and when an object has not been classified the object will be classified as restricted access policy
- The classification process is typically not a complicated process. Most security assessments take around one week – but more complicated ones may take up to 2 months.
- The national security legislation applies to authorities – but may "spill over" to private actors if the information object concerns areas of national defence. This can make cooperative ventures within data analysis a bit tricky.
- The best way to overcome this is to write a SUA (säkerhetsklassat upphandlingsavtal), which can be a compliment to a business and/or project agreement. Such an agreement can force all parties not to disclose any secrets – even after the project is completed.
- Through a SUA agreement STA can conduct security clearance of all personnel involved in the project. Such clearance may involve both interviews, references and register extracts.

13.3 Other considerations

- **Data-driven decision-making needs to prove itself:** While there is much interest within STA around data-driven approaches to decision making of different kinds (among these policy-oriented decisions), the truthfulness and correctness need to be established.
- **The Role of STA:** how shall STA cooperate around data sources and data sharing with external actors (haulage firms, Google, Waze, OEMs etc.)? This is still an emergent ecosystem where roles have yet to be established.
- **Historical data is seldom stored for analytical purposes:** this typically means that data is incomplete, may have been recalculated and contain erroneous values.
- **Data is often insufficiently described:** there is currently a lack of metadata – what the data contains, the quality of the data, and data sources may be reclassified and become unavailable.