Big Data voor Incident management

December 2014
Eindrapport

A project voor:

Corporate Informatie Voorziening (CIV)

Background

- In the last decades, many studies have been dedicated to the analysis of road traffic incidents and at the assessment of which factors affect the probability of incidents.
- Broadly speaking, the most common form of incident prediction model is expressed as an incident formula, where Incident(s) “I” at time “t” and place “p” are a function of traffic “T” at a set of risk conditions “c” at the same time and place: $I(t,p)=f(T_{t,p})*g(c_1^{t,p}, c_2^{t,p}, \ldots, c_n^{t,p})$.
- Incidents can be defined in various forms, ranging from a binary estimate (yes, no) to a probabilistic distribution in a certain area. Also, the variable t and p can be point or range values, such as incidents on a stretch of highways during a 1 day period.
- There is abundant literature and a vast range of specific formulations of the incident formula. In the vast majority of the cases they seek to optimize the model parameters and better capture the cause-effect relationship between traffic, risk conditions and incidents to produce realistic, localized predictions.
- In the recent past the growing availability of incident data, traffic data and other data sets (such as weather) characterized by long time series and very detailed time-space resolution has led scientists to consider alternative approaches. Instead of seeking to model the cause-effect relationship leading to incidents (and thus predicting incidents based on the onset of causes) they look at pure data-driven models.
- These models do not seek to “understand” the phenomena but to maximize the ability of sifting through massive amounts of data to pinpoint a certain occurrence of the phenomena without attempting to explain the cause-effect relationship.
- This “Big Data” approach to incident prediction is worth exploring: it complements traditional incident management practices and is a way to attach value to data assets and data capturing investments.

Project

- The project scope is to formulate a set of hypothesis to validate a Big Data approach to incident prediction through exploration of candidate datasets and of prediction methods.
- The project focuses on three datasets:
  - Registry of highway incidents
  - Traffic measurements of vehicle speed, flow and travel time
  - Weather data
  - The reference data spans the period 2006-2014. Samples of these data sets are used for the project and for exploratory data analysis.
- The project activities include:
  - Data Preparation, structure data cleaning and data, preparation of data sets for analysis and integration into Big Data testing infrastructure
  - EDA (Exploratory Data Analysis) and descriptive analysis of the data
  - Analysis of the spatial and temporal dimensions of incident data; identification and analysis of patterns of data and clusters;
  - Assessment of preliminary prediction results and formulation of hypothesis for implementation and further dedicated studies
- Simply stated, the project seeks to propose data-driven predictive models for the occurrence of highway incidents based on the spatio-temporal patterns discovered in the analysis of massive historical datasets of incidents, traffic and weather.

Structure of the analysis

Traffic dataset
- Detailed speed and intensity on the main roads
- Every 500m, aggregated 10mins
- From 2006 to 2014
- Implicitly includes causes and effects of incidents

Weather dataset
- Detailed time-space historical weather data: multiple parameters
- Short term weather prediction (hourly) Long term weather predictions (up to 14 days)

Incident dataset
- Individual incidents, categorized and described
- Includes incident information and incident management information
- Includes road status information and some context information

Exploratory data analysis (EDA)
- What are the characteristics of each data sets ?
- What are the time-space features?
- Which variations do we observe over time?

- Explore the context features of each data sets
- Assess the reasonable utilization of each data set and which parameters can be predicted
- Visualize the content
- Explore qualitative relationships between the three datasets

Hypothesis formulation
- Which prediction models appear as potentially applicable?
- Which predictions are plausible?

- Identify a small set (possibly three) candidate prediction models
- Perform a first analysis of the models and of the results
- Advise on how to proceed through a more formal and targeted modeling phase.

EDA FOR INCIDENT DATASET

Incident database

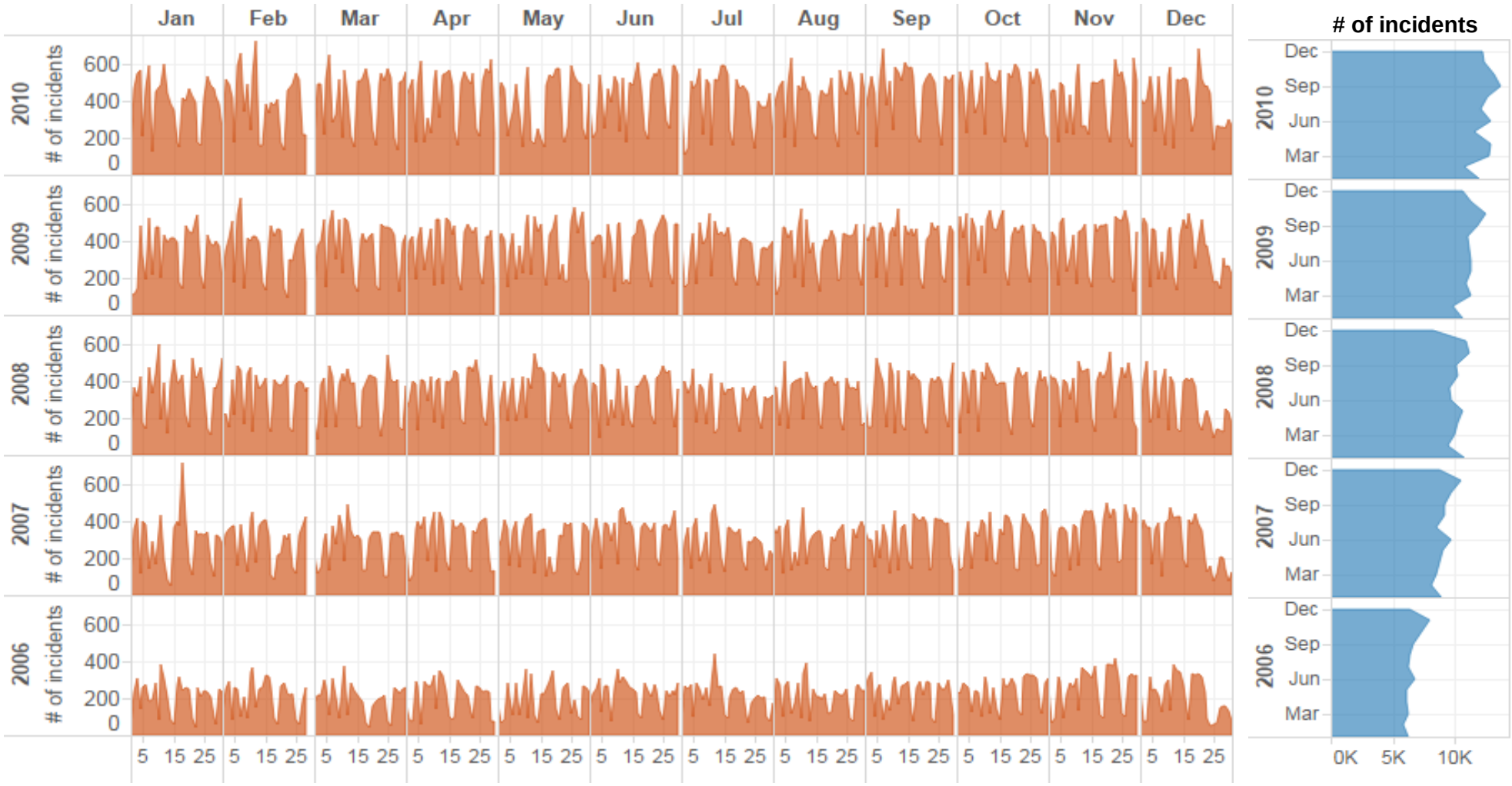
The dataset contains 590,737 incidents for the period of Jan 1, 2006 to Dec 31, 2010. Each incidents is described with 74 attributes, that provide information of several types, such as:

Attribute type	Example of attribute
Time	• Time of incident occurrence • Registration time • Time when did respective emergency services arrive • Duration (how quickly it was resolved) • etc.
Location	• incident coordinates • region, • road/highway name and number • hectometer point • indication of road direction • Road section • etc.
Administrative information	Indication of an authority, particular division and inspectors responsible for incident management
Incident descriptive characteristics	• classifications of incident types • type of involved vehicle
Incident attributes	• Injuries • involvement a road inspector • presence of evidences on the road • Etc.

Example of an incident record

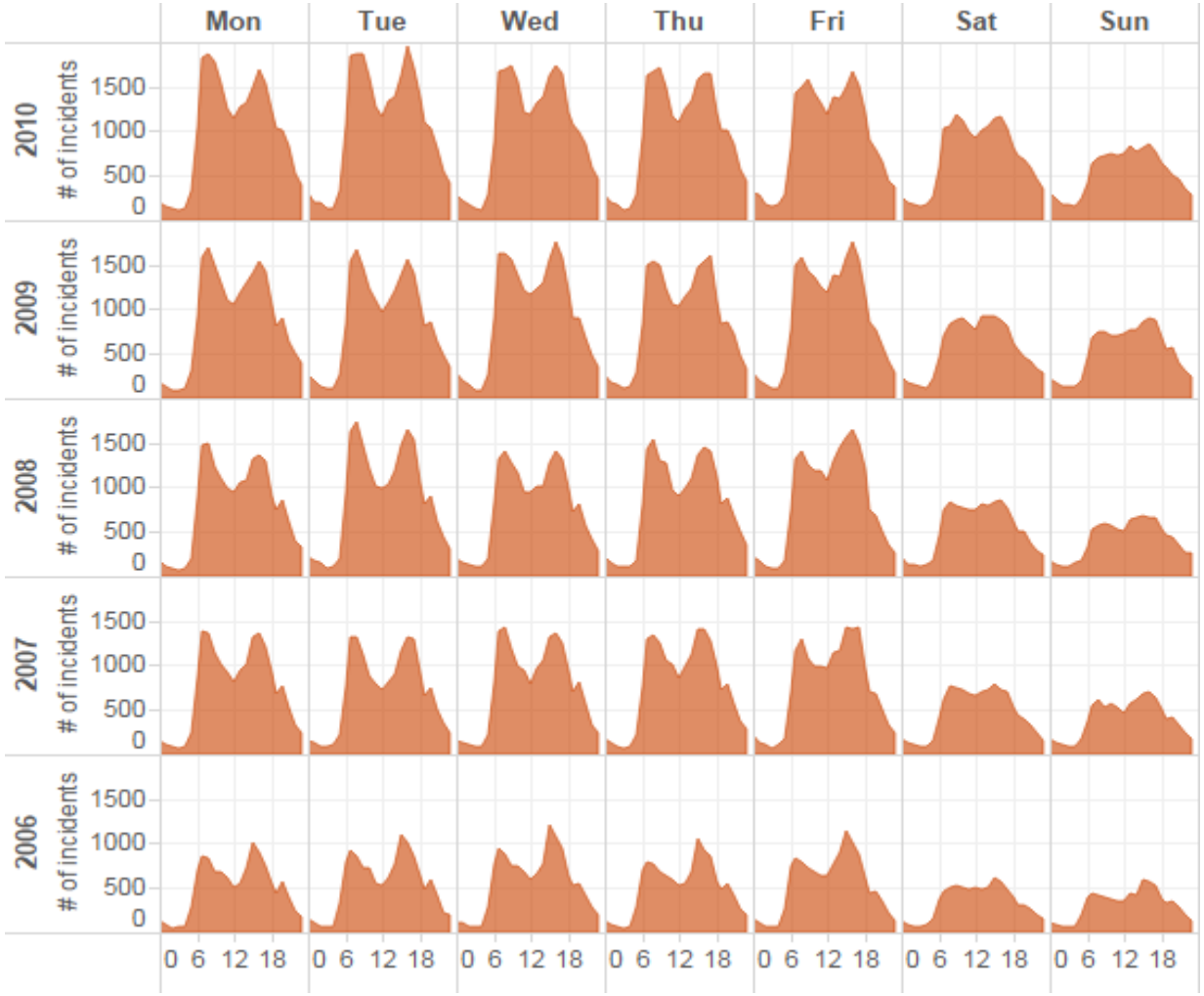
Attribute type	Exemplary values of main attributes
Time	• Beginning of an incident / notification time: **14-03-2006, 08:11:40** • Activation of a road inspector: **14-03-2006, 08:12:00** • Arrival on site: **14-03-2006, 08:31:00** • End of incident / departure of a road inspector: **14-03-2006, 08:59:00**
Location	• X, Y coordinates in RD new system: **133255, 453074** • Geographic region: **Utrecht** • Road name / number: **A12** • Hectometer point: **57.5** (57.5 km from the begging of the A12 highway) • Indication of a road direction: **Re** (right side part of the highway looking from The Hague) • Road section: **Oudenrijn – Galecopperbrug**
Administrative information	• Road authority: **RWS**, • Administrative division: **Verkeerscentrale Midden Nederland** • Administrative region: **Utrecht** • Name of responsible road inspector: ***
Incident descriptive characteristics	• Classifications of incident type CL1: **Accident** (ongeval) • Classifications of incident type CL 2: **Accident with injury** (ongeval met letsel) • Type of involved vehicle: **Passenger car** • Type of road: **Ring**
Incident attributes	• Injuries: **Yes** • Involvement a road inspector: **Yes** • Presence of evidences on the road: **No**

Variation of incidents over a year



The number of incidents does not show seasonal variations but we observe clear weekly patterns.
Furthermore, total number of incidents steadily increases in time, faster than the increase of vehicles on the road.

Variation of incidents over a week

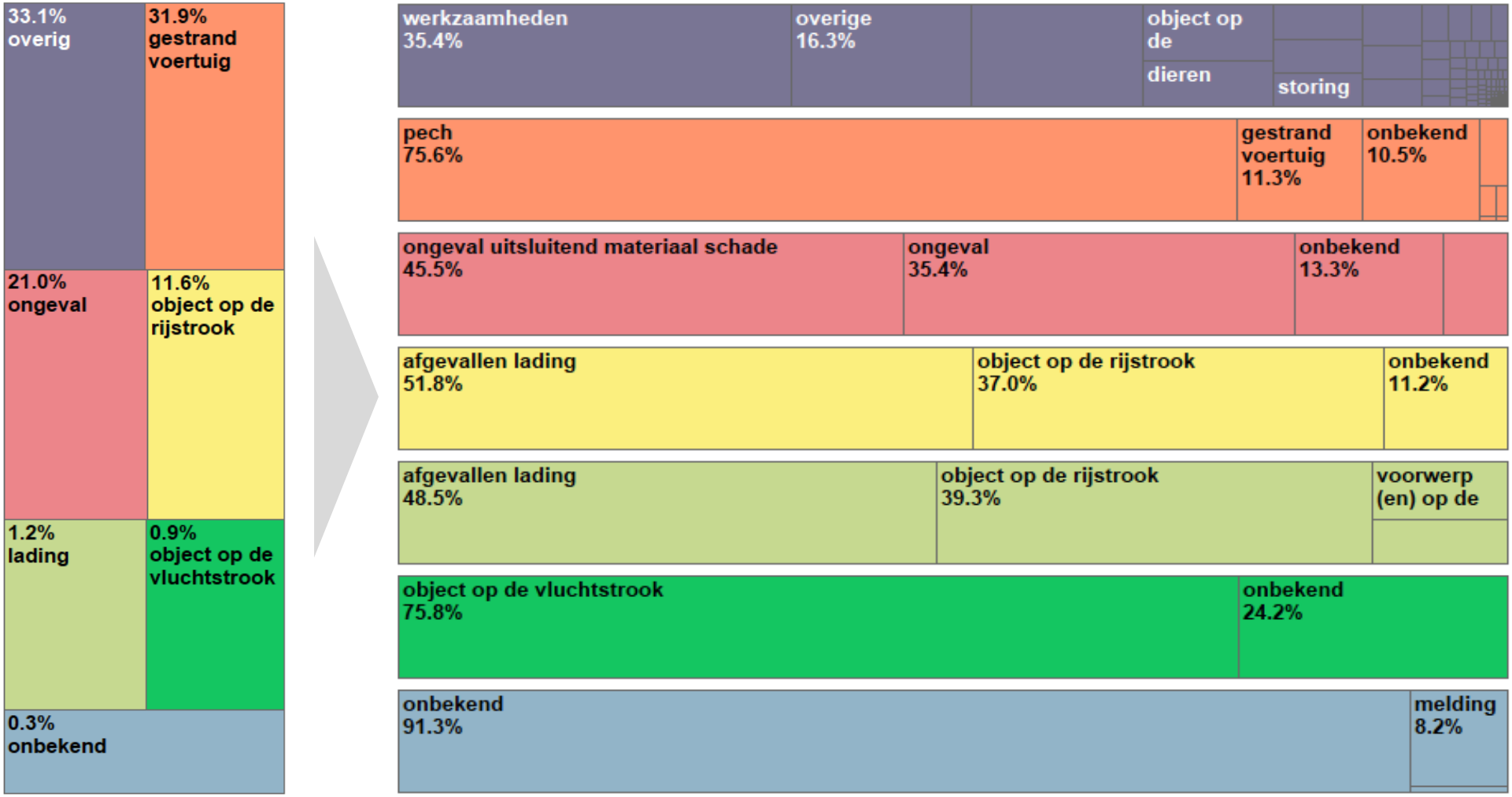


Distinct daily pattern with two peaks: around 8am and 5pm, corresponding to peak traffic times.

The # of incidents is much smaller over weekend and the peaks are less distinct.

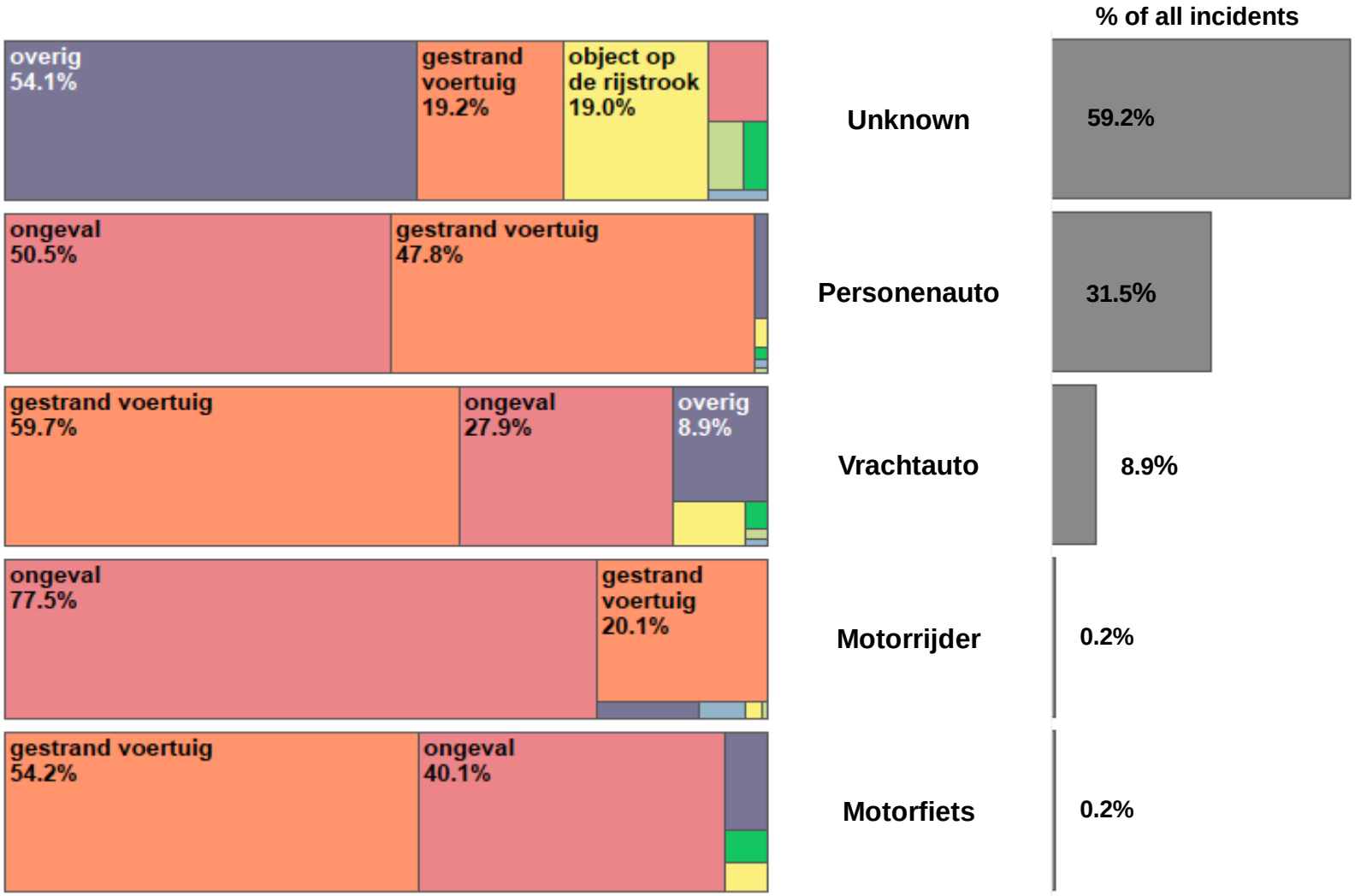
Categories of incidents (2006-2010)

Incident type CL1 (left) and categorization of each class based on CL2 data (right).



(note the logarithmic scale for the size of squares)

Incidents for different types of vehicles (2006-2010)



Spatial distribution of incidents (2006-2009)

overige

gestrand voertuig

ongeval

object op de rijstrook

lading

object op de vluchtstrook

onbekend

Spatial pattern is different for different categories of incidents e.g. “overige” along the a50, “gestrand voertuig” in Rotterdam and Amsterdam and Eindhoven. Categories “lading”, “object op de vulchtstrook” and “onbekand” are visible only on a limited part of the country. This may be due to a difference between the classifications of traffic centers, or data gaps.

EDA FOR TRAFFIC DATASET

Traffic dataset

There dataset contains traffic measurements at detection loops along the main Dutch roads and highways for the period 2010-2014.

The speed and # of vehicles are measured by around 14,000 loops, located every 500m along the roads and positioned at each lane for both directions of travel.

The data is available at a maximum resolution of 1 minute intervals: we use data aggregated at 10 minutes intervals. In total, almost 23,000 data records are produced every 10 minutes.

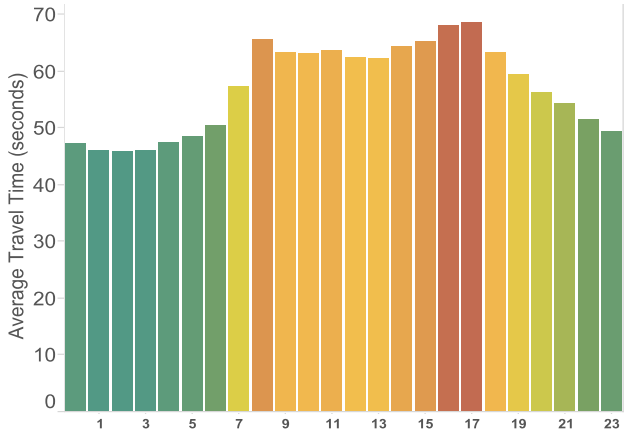
The dataset contains three traffic measures aggregated over 10 minutes:

- Avg. speed in km/h measured by a given induction loop
- Avg. flow of vehicles, i.e. avg. # of vehicles passing through an induction loop
- Avg. travel time for a road segment between two loops

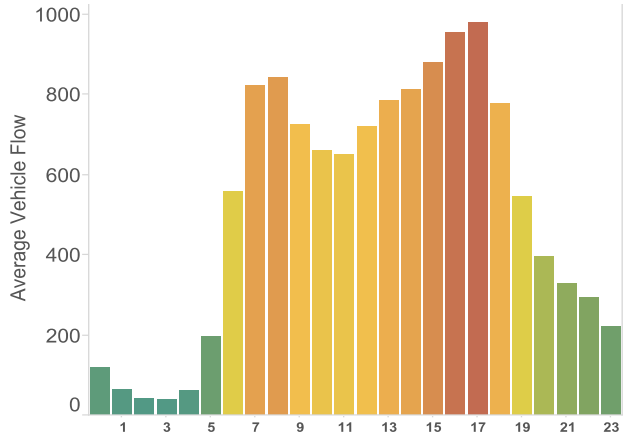
For ±20% of loops, the measurements are further categorized by the length of the vehicle.

The travel time is measured for around 4 K road segment, without distinction of lane or vehicles' length.

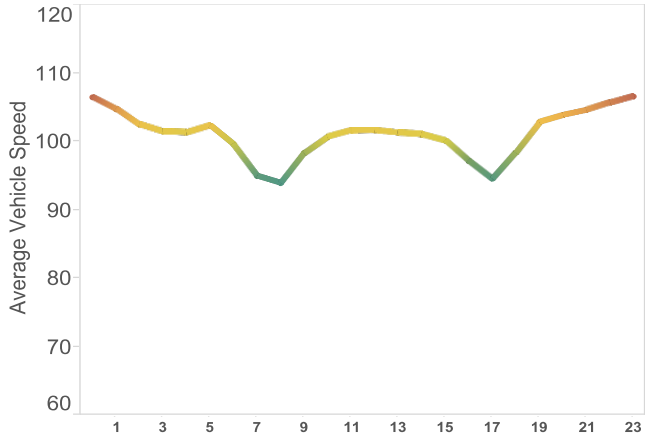
Three types of traffic measures



Travel time
Average travel time (24 h) between consecutive loops.

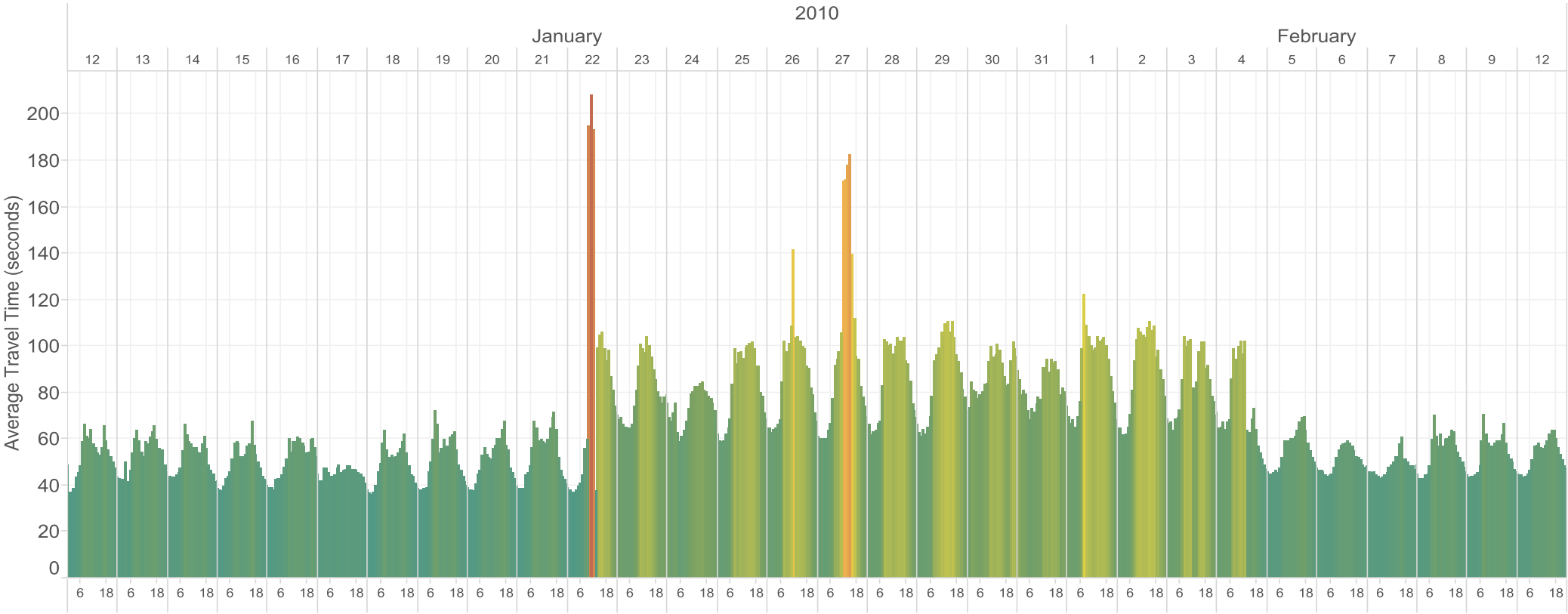


Flow density
Average vehicle flow (# of vehicles per hour)



Speed
Average speed per hour.

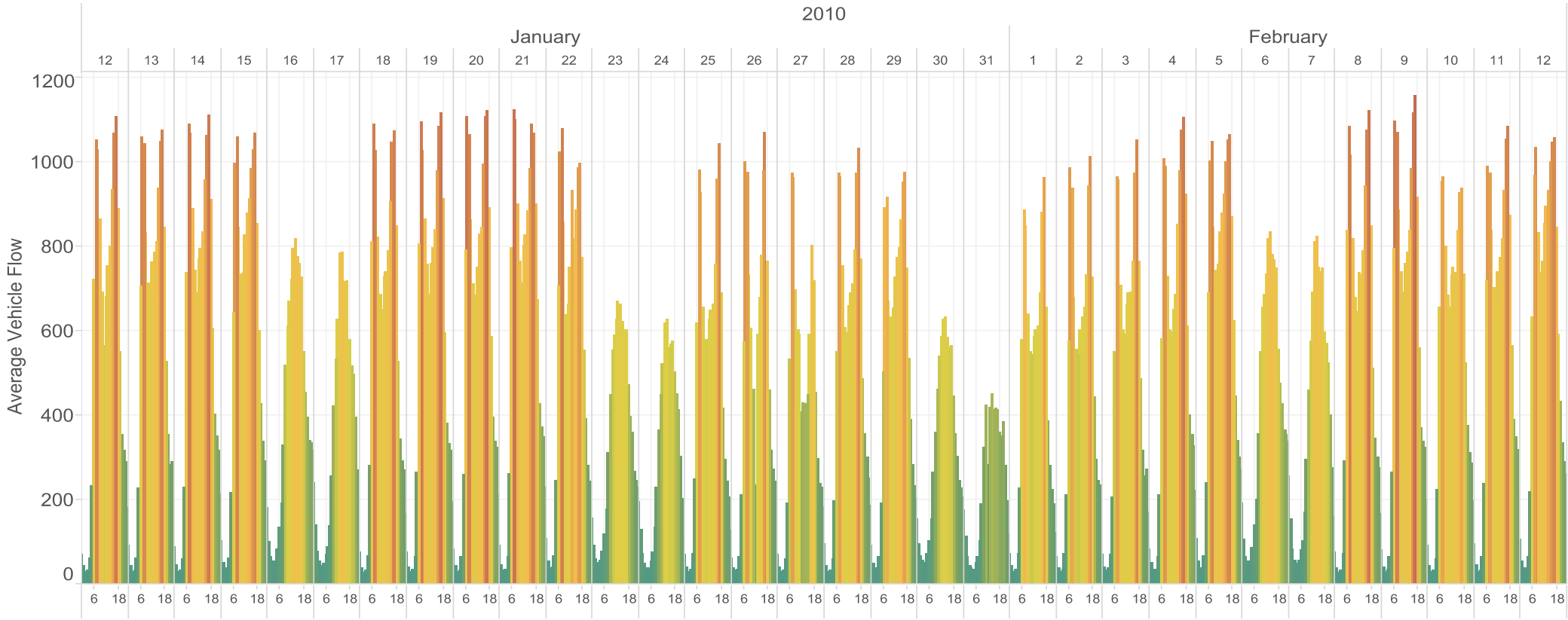
Fluctuations of travel time
(sample period of one month)



The travel time follows a is stable daily pattern, over weekdays and weekends.
The increased travel time between end of January and beginning of February – with the visible peaks of Jan 22 and Jan 27 - is connected to a period of extreme cold weather and snowfalls.

Fluctuations of vehicle flow
(sample period of one month)

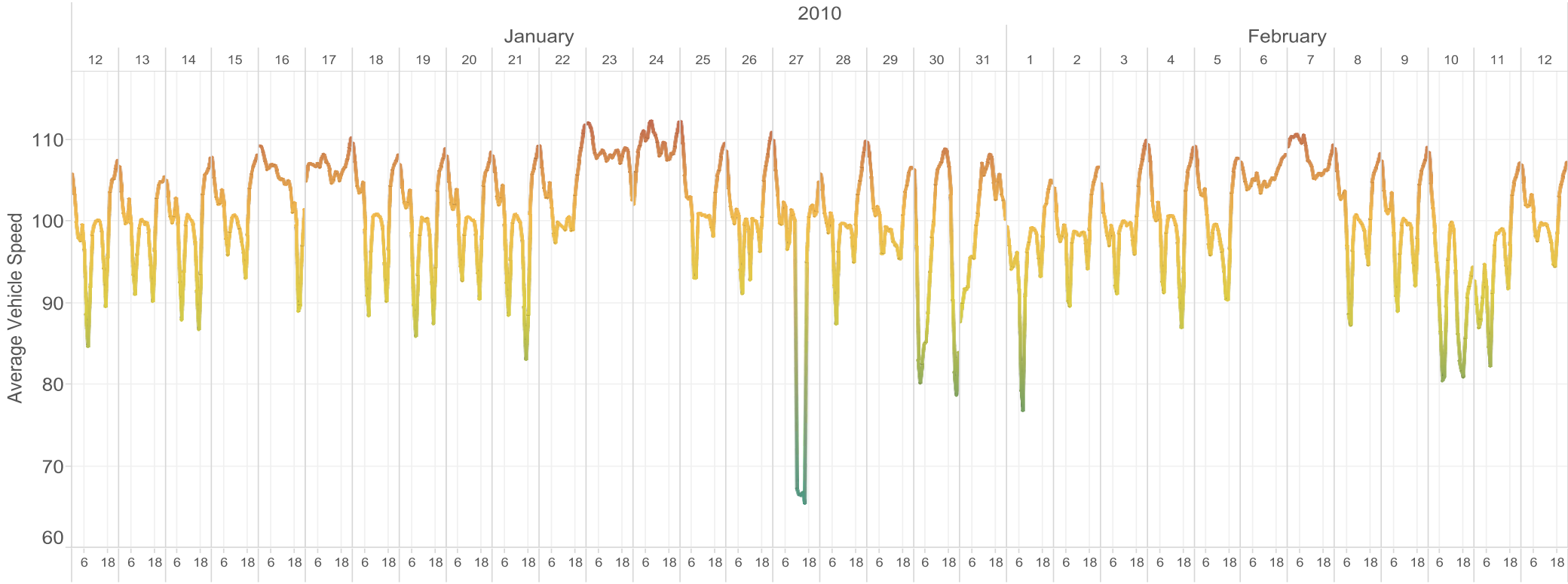
Vehicle Flow during 1 minute (averaged over an hour)



There is a stable daily and weekly patterns, with the # of vehicles significantly lower on weekends

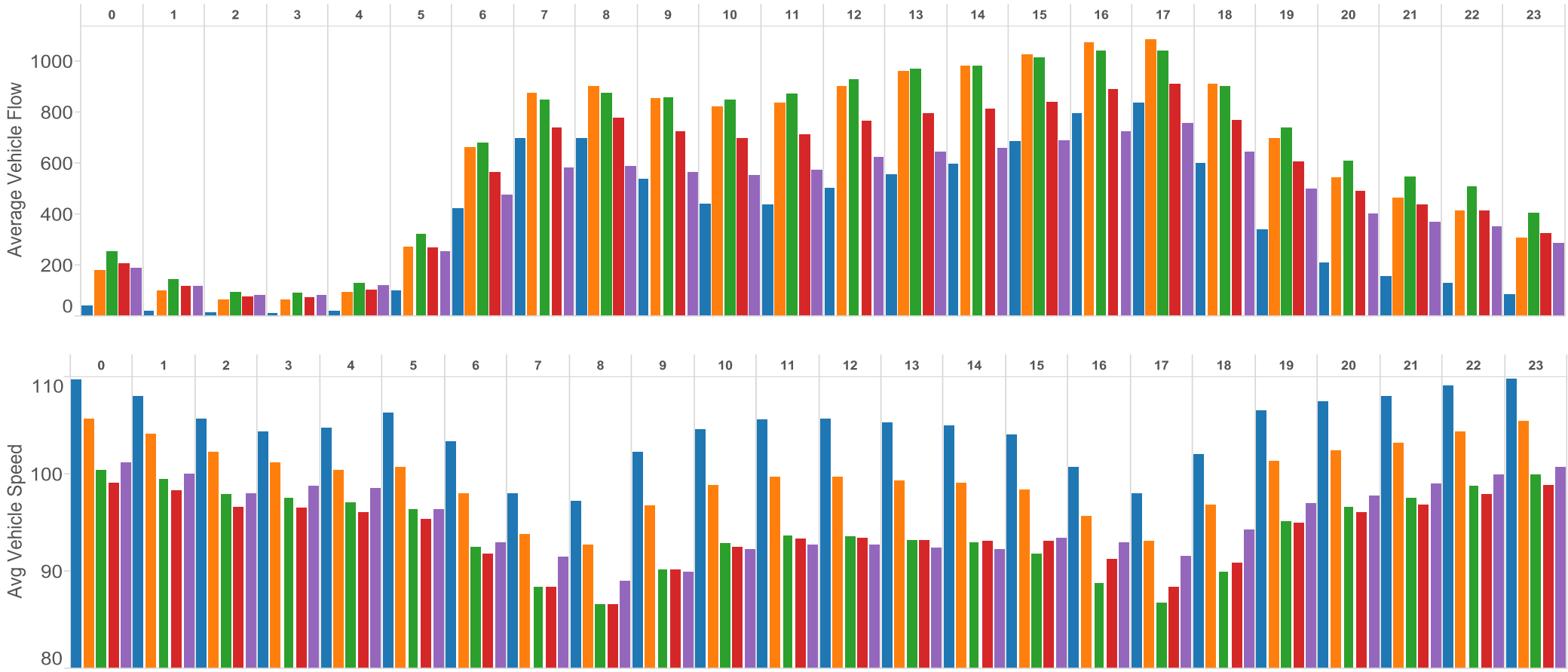
Fluctuation of vehicle speed

(sample period of one month)



The average speed (over one hour) is higher over night when the # of vehicles is lower.

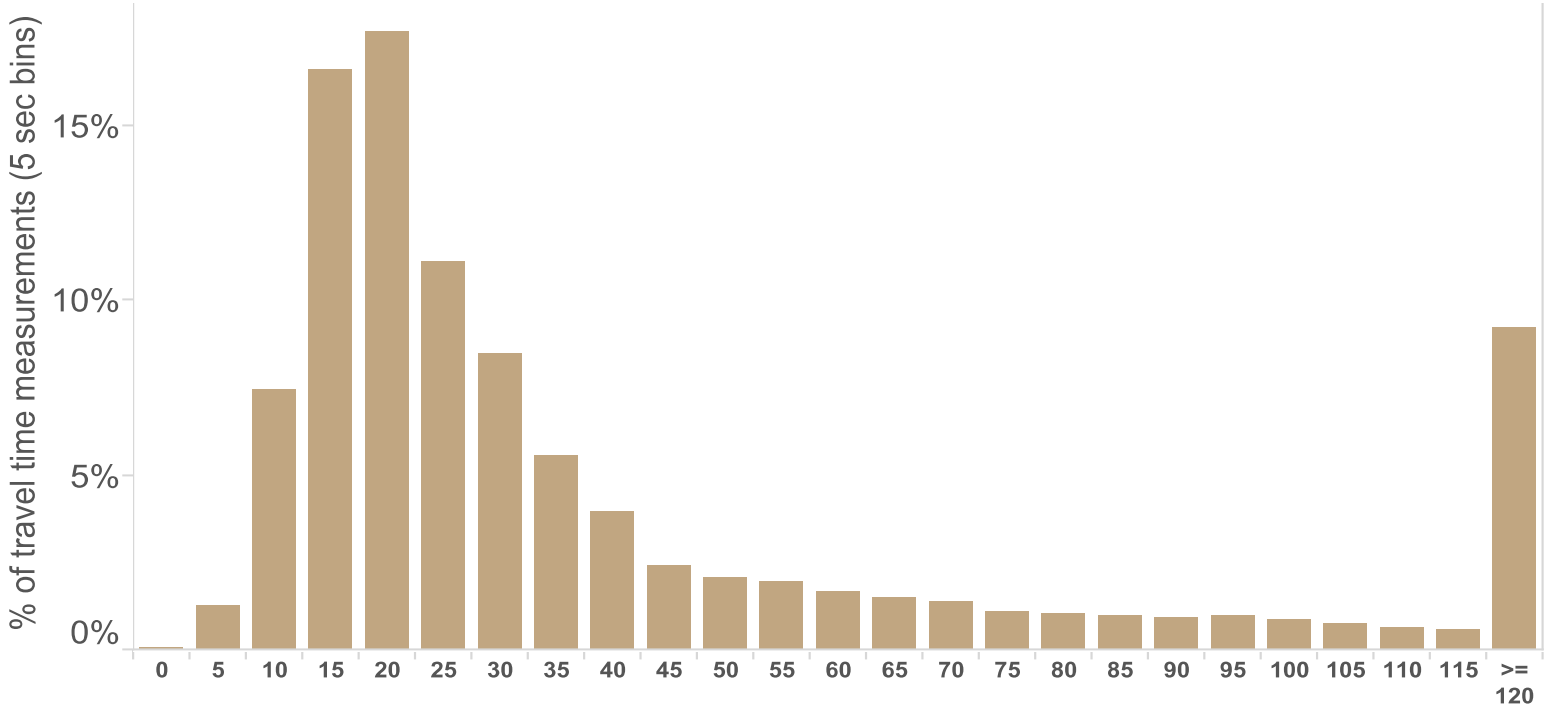
Avg. flow and speed on different lanes over 24h



Lane 1
Lane 2
Lane 3
Lane 4
Lane 5

Each direction of travel is divided into lanes. Numbering starts at 1 (far left lane). The number of the slow lane (rightmost lane) can be anything between 2 and 5.

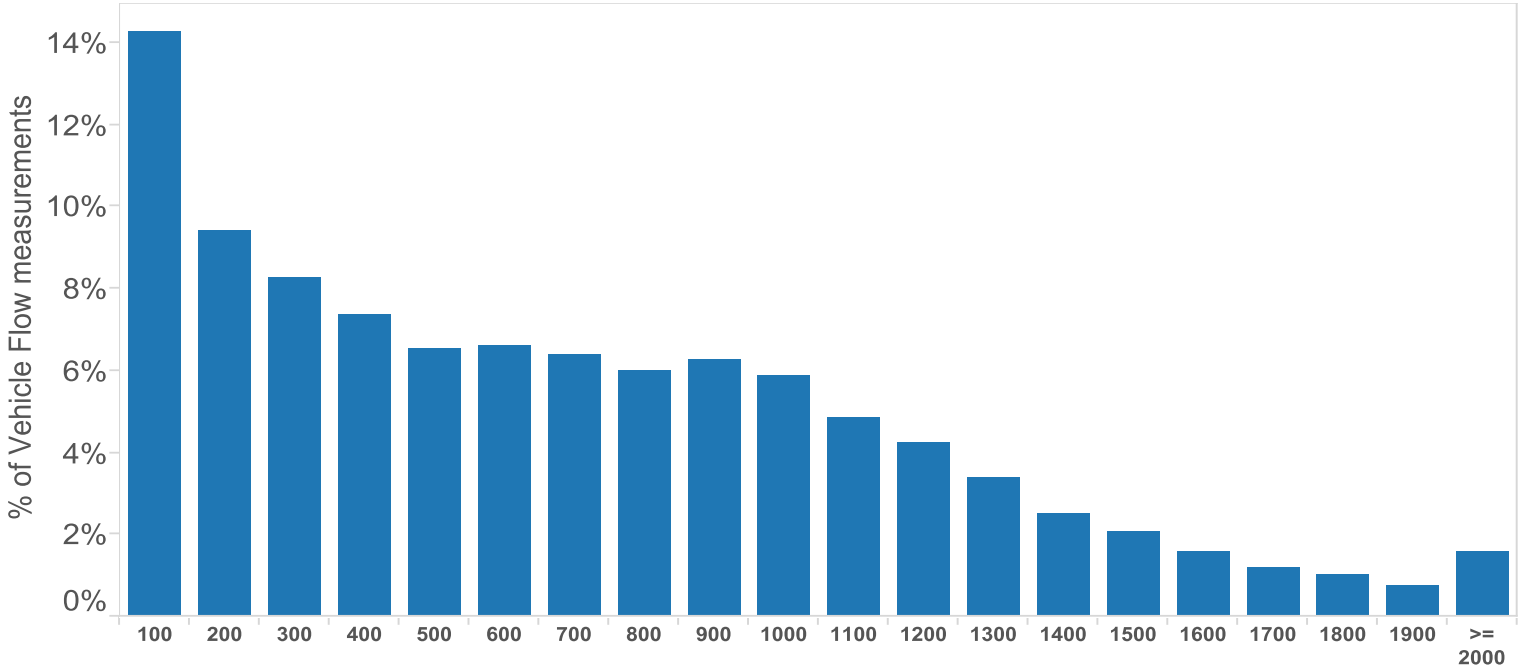
Distribution of travel time measurements between loops



Most of the recorded travel time fall within the period of 15-20 seconds (corresponding to a speed range of 90-100 km/h).

Around 10% of measurements indicate travel time values above 120 s, corresponding to traffic jams or other disruptions, including incidents.

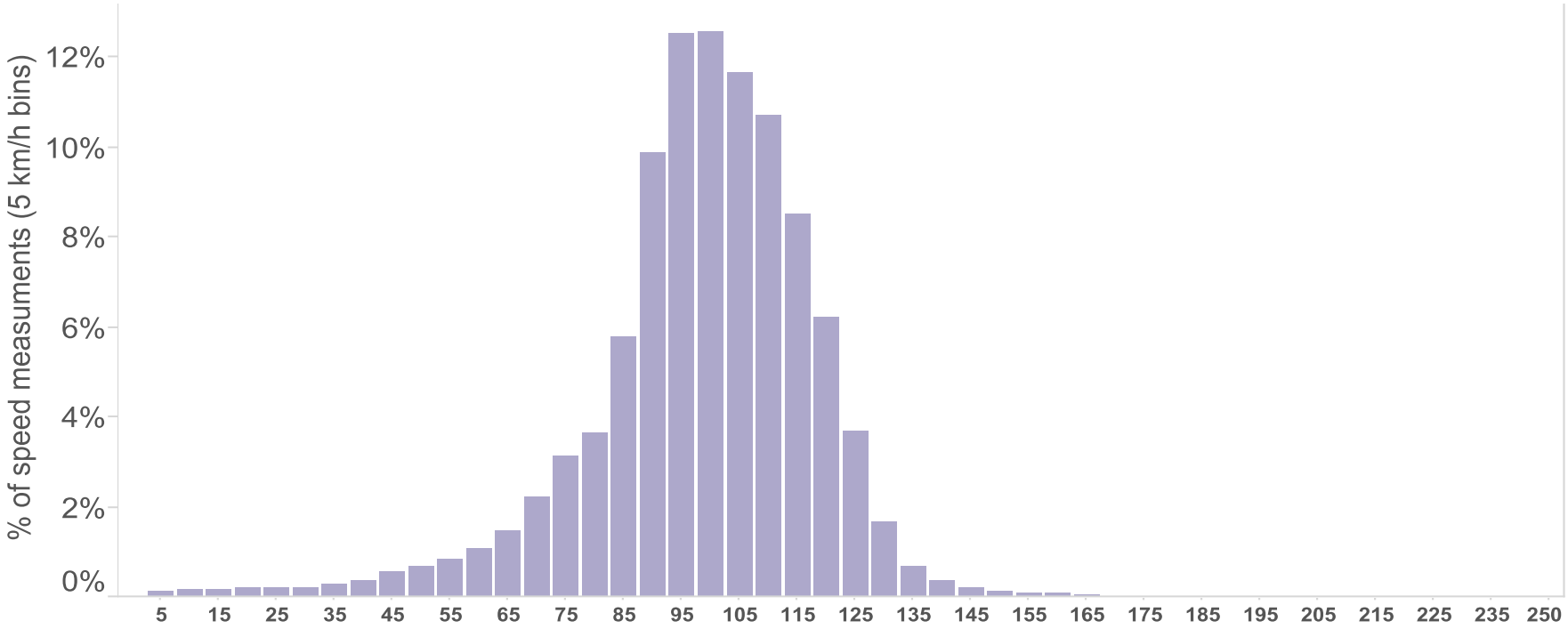
Distribution of flow measurements across all loops



14% of flow measurements register up to 100 vehicles during 1 minute (on average)

Distribution of speed measurements

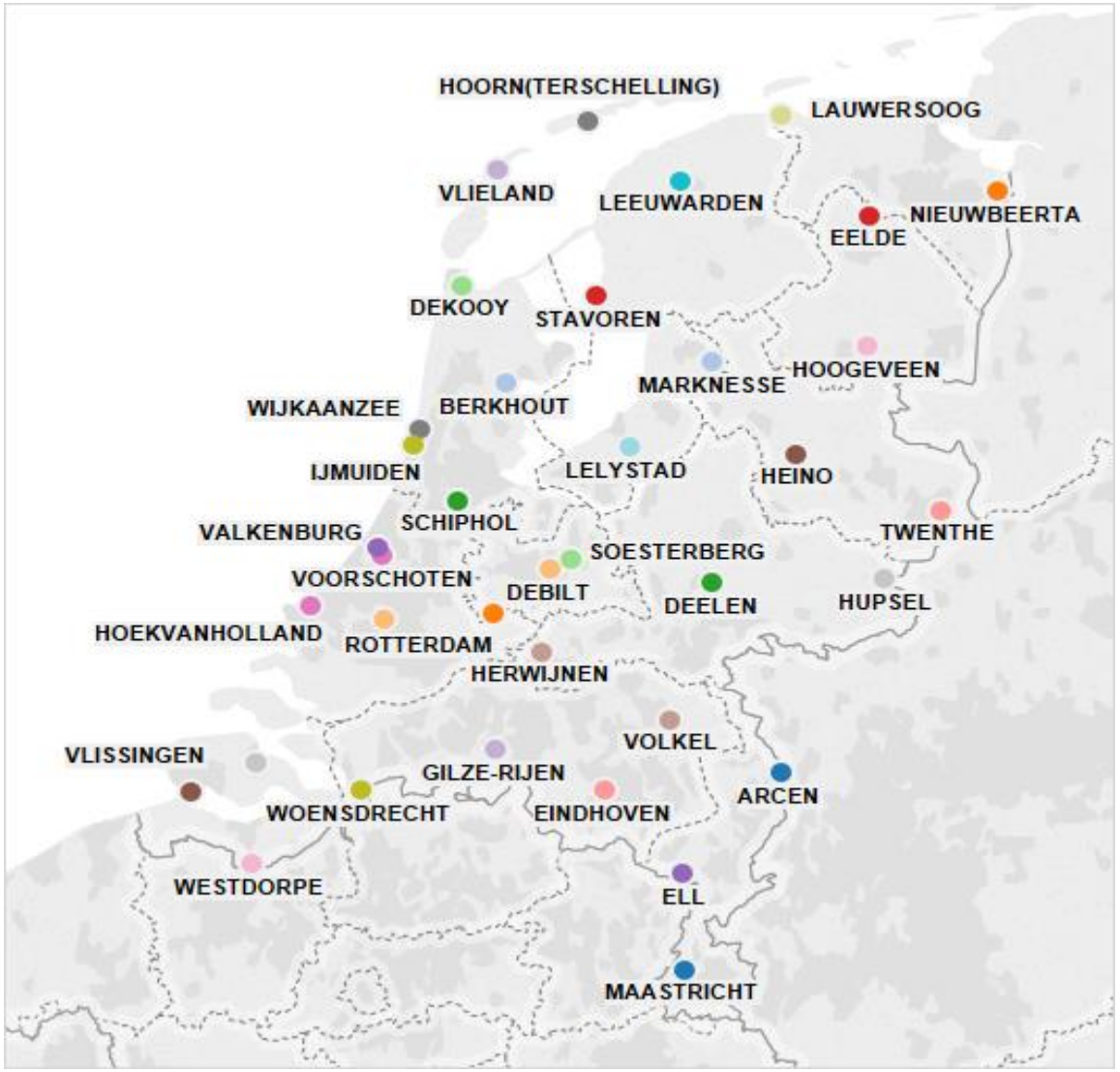
(sample period of one month)



Most speed measurements are centred around an avg. speed of 100 km/h (average across the entire network). Speeding is observed up to a maximum of 250 km/h.

EDA FOR WEATHER DATASET

Weather data



Daily measurements of:

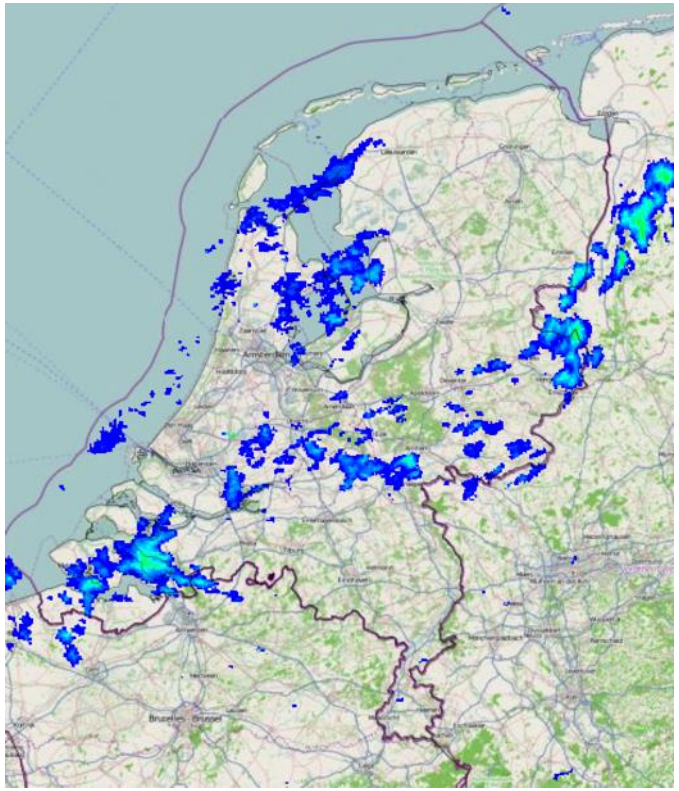
- temperature
- sunshine
- cloud cover and visibility
- air pressure
- wind
- precipitation

From automatic weather stations across the Netherlands

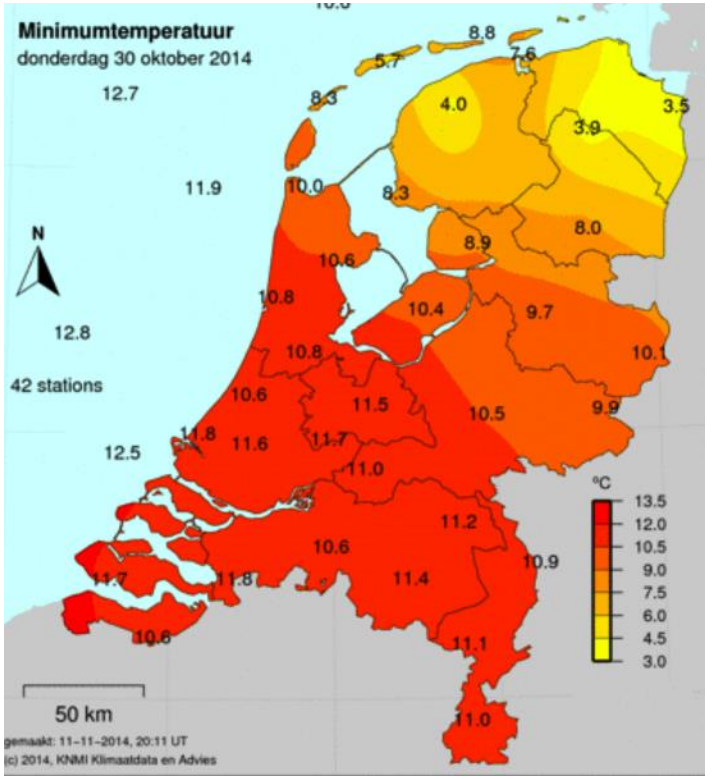
Available from Royal Netherlands Meteorological Institute (http://www.knmi.nl)

Weather data resolution

Weather historical data is available at high temporal and spatial resolution (direct measurements or interpolation) and can be associated to the road infrastructure precisely.

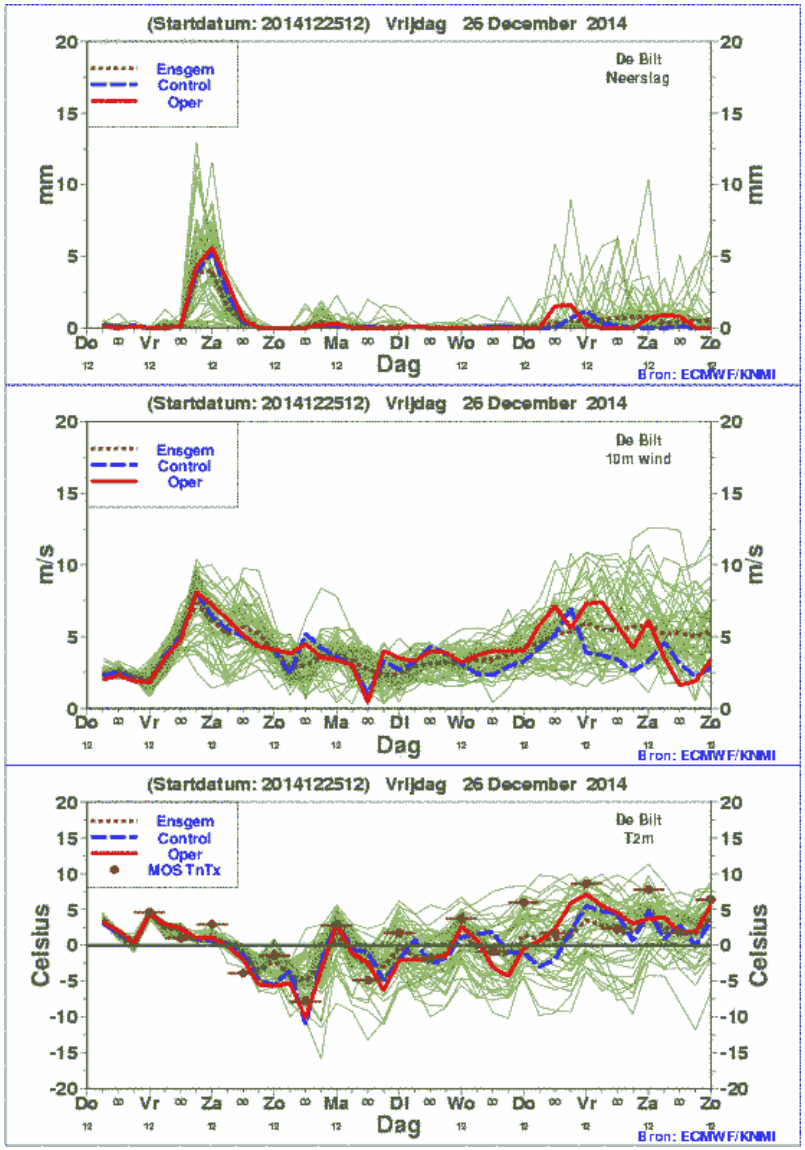


Precipitation data (29 October 2014, 8:25am)



Minimum temperature (30 October 2014)

Data forecasting



Several forecasts are available at country and local levels, for all main variables that affect the road condition and travel safety.

Models include indications of uncertainty. The data on the left shows “consensus” estimates for precipitation, wind speed and temperature and the underlying models producing individual estimates based on specific assumptions.

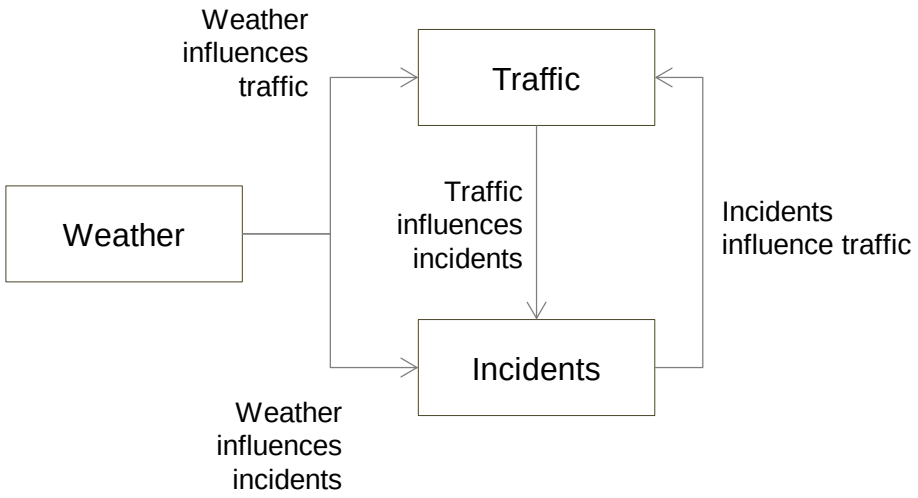
HYPOTHESIS FORMULATION

Data logics

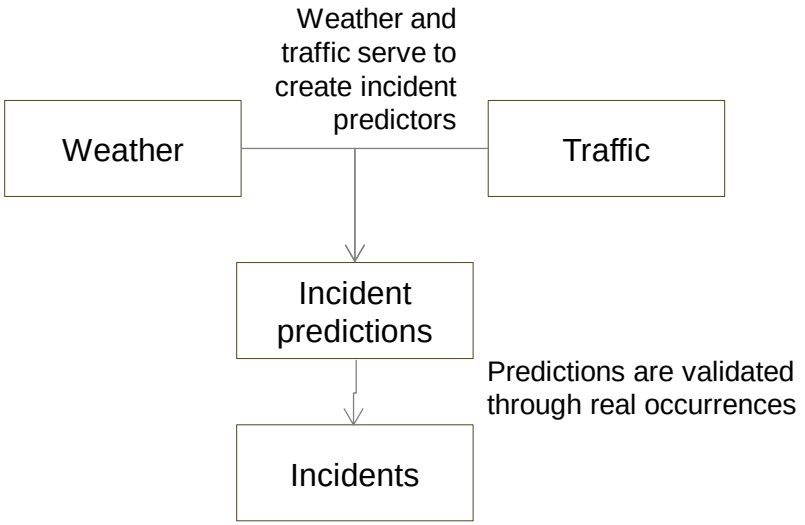
Predictions range from individual incident prediction to aggregated estimates (probability that at least one incident will occur in a certain area during a certain period of time).

The traffic data implicitly contains the effects of weather and incidents (in the form of travel disruption recorded through speed and flow alterations). By using weather separately in the prediction chain we seek to single out its influence on incidents.

The cause effect chain



The prediction chain



What's predicted?

There are multiple choices for what's predicted. The questions we need to answer are:

- What's the time horizon of prediction?
  - One hour?
  - One day?
  - Longer?
- What's the spatial dimension?
  - A point on the road network?
  - An arbitrary section (e.g. 500m)?
  - A logical section (e.g. between intersections)?
  - A specific hot spot area (e.g. an intersection area)?
- What's predicted?
  - Any incident?
  - Incidents with high traffic impact?
  - Multiple occurrences of incidents?

Examples:

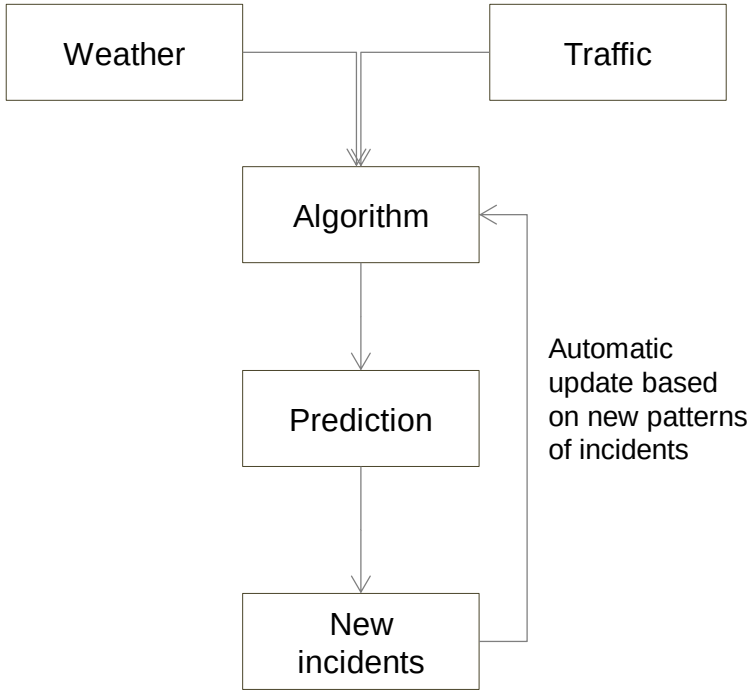
- Probability of an incident to occur in the next hour in a section of 500m.
- Highway sections with the highest chance of incidents in the next 4 hours
- Location of the next incident, within 60 minutes
- Hottest incident section in the country, in the next 24 hours
- Chance of a major incident in the road network in the next 2 hours

Methodologies

The large amount of traffic, weather and incident data allows us to experiment with purely data-driven methodologies for incident prediction, without making modeling assumptions at the outset. The goal is to verify their suitability for various prediction outcomes as well as their ability to adaptively incorporate new data and adjust predictions without human intervention. The starting point is to utilize the following methods:

- Logistic regression, which is naturally suited to the prediction of probabilities for incidents.
- Artificial neural networks, particularly in the deep learning setup, with the aim of discerning correspondences between patterns of weather, traffic and incidents.
- Experts systems of online learning, in which the prediction mechanism is continually adjusted based on incoming data.

Adaptive incident prediction



COPYRIGHT ©
COLLECTIVE SENSING AT UNIVERSITY OF SALZUBURG AND CS RESEARCH FOUNDATION AND/OR THEIR AFFILIATES
ALL RIGHTS RESERVED