CSRESEARCH

Incident prediction with neural networks
Voorspelling vanuit een anticiperende overheid

REPORT

A project for:

Rijkswaterstaat
Ministerie van Verkeer en Waterstaat

CSRESEARCH

Contents

1

BACKGROUND

Project background

The large number of road users on the crowded Dutch roads leads to significant congestion and incidents. On an annual basis there are more than **100,000** incidents and traffic disturbances, approximately **270** every day (from small events to large traffic accidents).

Incident Management (IM) is one of the key management practices to address incidents. The primary goals in IM are to ensure the safety of the roads and to restore the vehicles flow after an incident. More specifically, IM aims at:

- Increasing the responsiveness of incident handling, including securing the incident site;
- Improve the control or redirection of the flows;
- Reduce the consequences for incident victims;
- Streamline the cooperation between emergency response actors.

The visibility on risk sections of the road highway, or on circumstances that may lead to incidents and traffic disturbances, is of general interest for incident management. The ability to predict when and where an event will happen would have a major influence on the ability of mitigating the causes and respond appropriately.

Given the multitude of factors that lead to incidents and traffic disturbances this is probably impossible. However, improving the ability of predicting, even marginally, could be of significant value. The increasing availability of data (incidents data, traffic, weather, other data) opens the possibility of exploring methods of Artificial Intelligence for incident prediction.

Project rationale

Traffic flow can be affected by a multitude of events, ranging from minor local disturbances (an object on the road) to major incidents blocking the entire traffic flow for hours. In line with the practice of RWS, we call ”incident” any travel flow disturbance that leads to an action by road authorities to regulate and/or re-establish the flow.

Some of these incidents are the result of random external causes (a branch of a tree falling on the road). Many others are influenced by known factors (speed, vehicle density, weather etc.) although the influence appears as complex and highly non-linear.

While it is probably impossible to precisely predict incidents for a specific time and place, it is realistic to expect that the interpretation of patters of the factors that influence incidents can lead to better information compared to historical data alone.

The rationale for looking at incident prediction is twofold:

- To provide foresights that lead to better resource allocation
- To provide foresights that lead to more effective incident prevention measures

Incidents

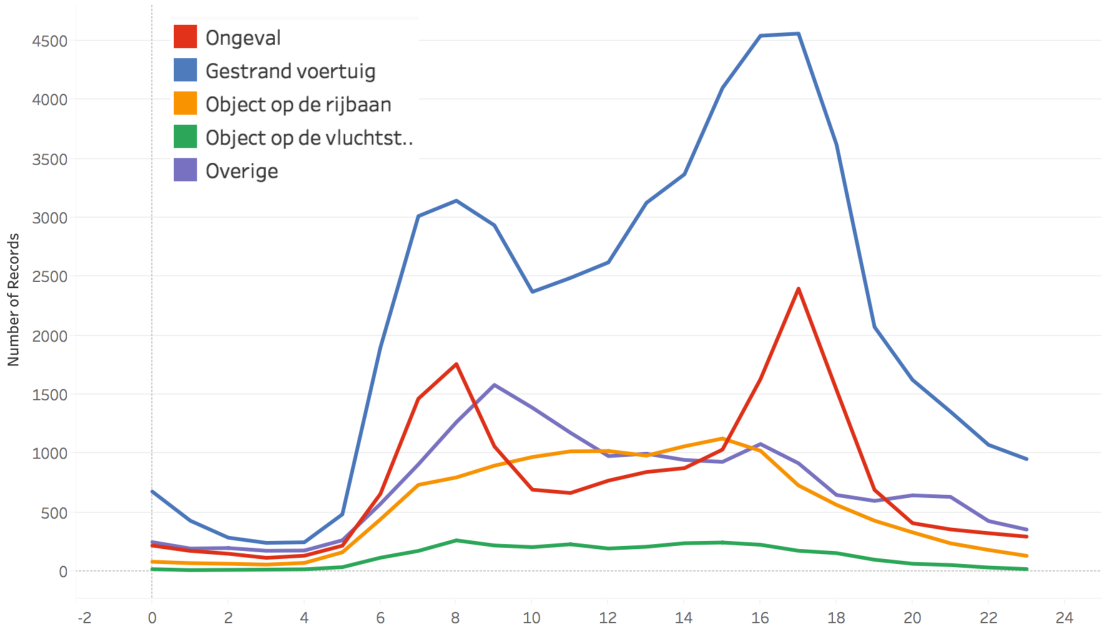
Incidents include a wide range of events that may impact the traffic flow. In the practice of road inspection and incident management, incidents are recorded in classes, as follows:

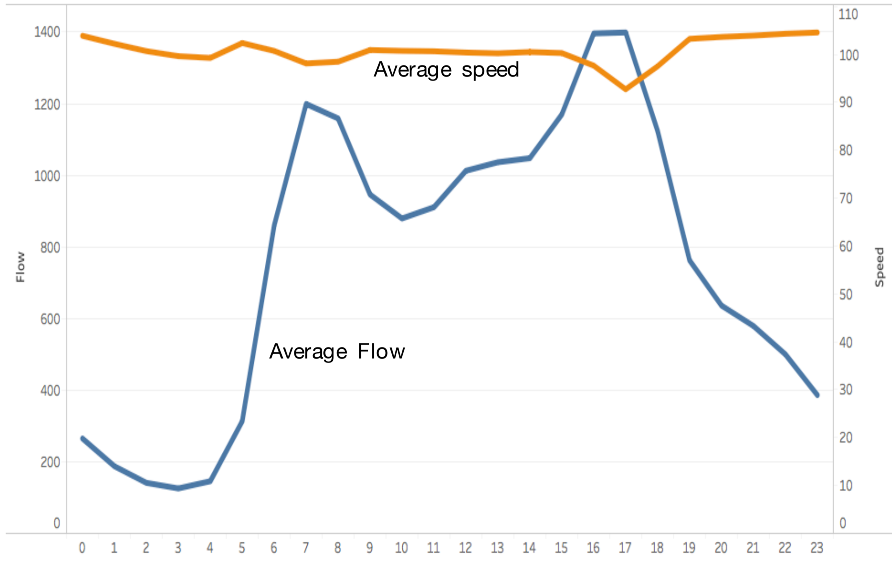
- Vehicle problems (*Gestrand Voertuig*): Incidents associated to a car breakdown such as engine failure, flat tire, etc.
- Accident (*ongeval*): accidents involving one or more vehicles, such as cars, busses, trucks, leading to material damage, injuries or fatalities
- Objects on the road: incidents related to the presence of objects, such as wood, rubber or other objects.
- Other: other circumstance leading to a traffic impact, ranging from oil slick, to animals or temporary road activities.

For simplicity, we group the last two classes into “Other incidents”.

Incident data is recorded by multiple sources (Police, Towing Companies, Road Inspectors). We use the data provided by Rijkswaterstaat as a basis and review alternative sources as well.

Incident patterns





Ongeval and *Gestrand voertuig* follow a general pattern related to time and traffic flow.

The other incidents are more uniformly spread over the work day and the evening. This is probably related to the registration practice as well.

Year 2014

Local incident density in the 24 hours

Average number of incidents per hour per sections of 20km, year 2014

A12

Incidenttype	Hectometer..	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Ongeval	0					2	1	4	14	21	16	5	5	5	6	8	7	9	25	12	7	6	2	4	3
	20	3	3	2			1	6	19	17	18	4	4	2	4	9	12	22	43	38	11	3	3	1	4
	40	1	2	5	1	2	5	2	28	24	18	10	11	13	11	5	14	27	39	15	10	7	8	1	8
	60		2	2	3			8	14	17	12	11	7	6	10	3	11	14	23	16	7	3	4	7	3
	80	1	2	2	2	1		1	4	1		2		2	2	1		2		2	1	2			
	100	1	1	3		2	3	7	13	23	8	9	4	6	6	1	9	20	21	22	6	1	3	1	1
	120	2	1	1		2	1	10	10	27	7	16	14	6	8	6	12	16	30	19	4	2	7	1	
	140		1				1	1	5	10	2	2	4	7	9	5	2	3	5	1	2	1	1	2	1
Gestrand voertuig	0	13	5	6	4	3	17	20	35	41	36	24	32	37	50	56	52	52	50	53	42	28	18	17	18
	20	14	6	5	2	1	4	16	43	30	42	19	25	41	32	39	33	41	59	47	36	21	15	15	13
	40	11	7	2	7	3	7	27	24	36	33	32	26	29	24	34	35	43	43	38	29	15	18	16	17
	60	5	9	4		4	2	24	31	31	22	24	23	22	17	27	31	32	30	34	24	9	17	9	9
	80	1	2			1	1	11	8	7	2	3	7	6	5	6	6	6	8	6	7	1	1	2	1
	100	4	4		3	2	4	11	20	15	14	20	20	24	22	26	25	31	34	34	16	9	10	3	3
	120	4		1	2		8	28	30	28	35	36	34	34	33	49	70	51	72	51	16	16	13	11	9
	140	1				1	3	14	10	7	10	4	6	9	9	11	10	18	16	9	6	11	6	2	

A10

Incidenttype	Hectometer..	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Ongeval	0	8	6	3	2	2	3	9	27	44	30	19	21	14	16	23	18	21	63	49	21	11	9	12	9
	20	2	2	1	4		3	5	13	32	22	9	17	9	16	21	19	23	26	29	11	8	3	2	7
Gestrand voertuig	0	25	11	7	10	6	10	31	67	51	63	56	47	69	83	62	92	108	83	73	46	40	37	27	30
	20	16	7	10	9	2	6	14	32	35	47	34	40	37	46	46	79	46	59	41	34	24	24	19	20

A4

Incidenttype	Hectometer..	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Ongeval	0	7	2	2	3	1	1	12	15	34	24	9	9	10	17	5	8	14	18	21	11	7	9	5	5
	20	4	3	3	1	2	1	13	25	23	23	6	11	10	8	8	7	17	16	18	7			4	4
	40	4	1	2		1	3	7	24	21	20	10	5	10	8	6	12	27	35	15	13	2	5	3	3
	60	2		2				11	15	13	5	3	2	4	4	9	4	11	11	12			2		1
	200																						1		
	220	1					1		4	10	5	2	5	3	10	1	3	5	8	3	5	2	1	2	4
	240	1	1	1			3	1		2	5				1	2	2	8	9		3	1		2	
Gestrand voertuig	0	17	14	16	7	4	10	20	49	67	57	49	47	51	37	48	79	86	69	51	51	23	29	25	32
	20	19	9	8	4	4	8	26	41	44	35	39	33	34	39	38	53	64	61	61	36	25	29	20	27
	40	12	8	2	1	2	2	18	36	36	43	21	37	38	47	34	45	73	62	56	37	23	23	16	16
	60	8	7	3	2	5	4	21	30	28	23	38	24	22	21	30	46	35	37	38	21	22	6	18	5
	200									1		3			1		1			1					1
	220	1	2	2		3	2	4	12	9	9	9	11	1	7	9	10	13	8	8	7	6	6	5	2
	240	4		3	1		1	6	3	8	2	3	5	4	7	6	5	10	10	9	4	1	4	3	1

Hotspots

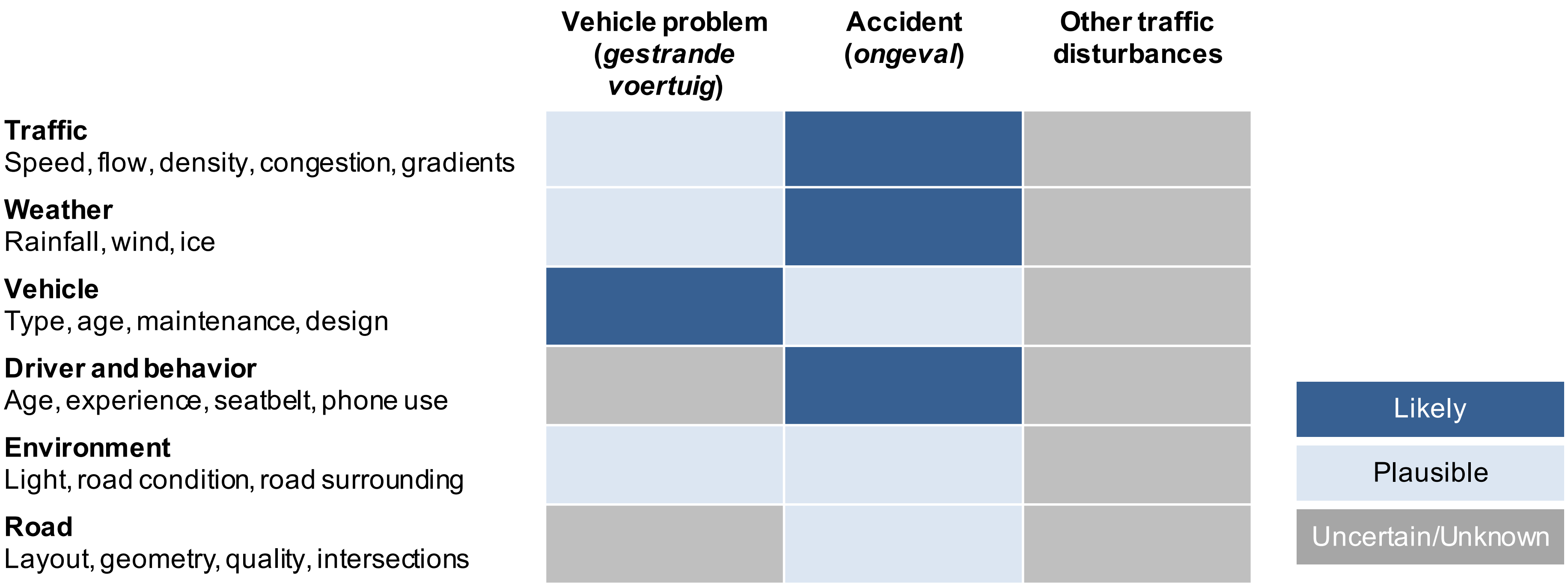
Average number of incidents per hour per sections of 5km, A12, year 2014

		Tijdstart																							
Incidenttype	Hectometer..	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Ongeval	0							2		5	8	1			2	1	1		5	3				1	1
	5								3	11	5	4	3	4	1	4	1	6	8	5	4	2	1	1	1
	10					1			7	5	1		1		1		3		10	4	1	1	1	2	1
	15					1	1	2	4		2		1	1	2	3	2	3	2		2	3			
	20	1	1				1	2			2	1	2		2		2	3	5				1		
	25	1						1	8	8	11	2	1		1	3	2	11	29	30	4	1	1		2
	30	1						2	7	5	4	1		2	1	2	6	6	3	3	3		1		2
	35		2	2				1	4	4	1		1			4	2	2	6	5	4	2		1	
	40		1	1			4	1	11	4	3	1		5	5		3	5	6	3	2	3	2		1
	45			2		1	1		2	1	4	2		3		2		1	6	2			2	1	1
Test area	50			1		1		1	1	3	1	1	2	1	4		1	4	3	2	2	1	1		2
	55	1	1	1	1				14	16	10	6	9	4	2	3	10	17	24	8	6	3	3		4
	60			1	1			5	12	10	10	9	7	3	8	2	7	8	17	11	5	2	1	6	2
	65				2			2		2		1		2			1	3	4	3				1	
	70			1					1	4	2	1		1	1	1	1	2	2		1				1
	75		2					1	1	1					1		2	1		2	1	1	3		
	80			1	2			1	3			1			1	1		1				2			
	85	1	1	1		1			1			1		1	1					1	1				
	90		1							1				1				1		1					
	95																								
	100						1			2					1		1	1	2	2			1		
	105			2			1	1	2	4		2	2	1	2		4	6	9	12	4	1	1		1
	110								2	3	1	2	1	2			1	6	2	5	1			1	
	115	1	1	1		2	1	6	9	14	7	5	1	3	3	1	3	7	8	3	1		1		
	120	2		1			1	4	6	12	1	4	4	1	3	2	5	8	9	3	2	1	2		
	125					1		3	1	6	2	3	2	1	2	2	3	2	2	4		1	1	1	
	130		1			1		3	2	4	2	6	5	1	2	1	3	2	10	7	2		2		
	135								1	5	2	3	3	3	1	1	1	4	9	5			2		
	140		1				1	1	2	6	2	2	2	5	6	3		3	3	1	2	1	1	2	1
	145								3	4			2	2	3	2	2		2						

Incidents and influencing factors: a qualitative perspective

Literature and expert knowledge indicates the influencing factors that are likely or plausibly associated to incidents. In principle, if there is a likely or plausible relationship between an influencing factor and an incident type, then it is reasonable to assume that it is at least theoretically possible to identify the onset of influencing factors leading to an incident, thus predicting.

In this project we focus on the effect of traffic conditions on incidents, specifically accidents (*ongevallen*).



Formulating the prediction challenge

Within the broad range of incidents, it is reasonable to address the challenge in concentric sets, asking the following questions:

- Is it possible/meaningful to predict (better than historical averages) any type of incident?
- Is it possible/meaningful to predict (better than historical averages) specific classes of incidents that have a known qualitative dependence on traffic and context factors?
- Is it possible/meaningful to predict (better than historical averages) the specific accidents that lead to material damages and/or casualties ? (ongevallen.)

There are two main strategies to address the challenge:

- Define an explicit **model** of incidents based on the influencing factors for which there is data available and optimize the model based on past evidence of incidents
- Use **machine learning** to learn - from past data - the relationship between influencing factors and incidents, without assuming a model a priori

Given the know modeling difficulties described in literature and the recent development of machine learning tools, there is a potential opportunity for address the questions above through machine learning approaches. The approach selected in this project is Deep Learning Neural Networks (NNs).

Summary of previous results

- RWS has commissioned two short exploration projects in 2015 and early 2016.
- The projects used various types of Neural Network and trained them over traffic and weather data on a test section of the A12 highway.
- The traffic/weather data is used to construct traffic “images” (sets of traffic and weather measurements) which are tagged with a binary label (incident/no incident) depending on the occurrence of an incident in an adjacent future time window.
- The studies indicate that significant correlations between loop data and traffic incidents can be detected by Machine Learning algorithms. At an aggregated level, the studies indicate that the network can detect patterns of traffic that increase the probability of an incident.
- Within the test framework (one road section) the results indicate that the algorithms have different performances for different prediction parameters (e.g. length of the road, duration of prediction).
- The studies also suggest that some classes of NNs perform better than others for some specific training situations, while some fail to properly train on the limited dataset available.
- The results, and the modeling exercise, are however severely affected by data gaps and data inconsistencies. The exact impact of this is difficult to measure, but it is likely crucial.

Suggested next steps from previous projects

The previous exploration projects suggest a number of possible avenues for developing a better perspective on incident prediction. The main suggested areas of development are:

- **Data modeling.** Addressing the extent and impact of data gaps, and the specific data gaps that affect the modeling. Testing ways to resolve, bypass or mitigate the problem.
- **Prediction strategy.** Establishing a perspective on which type of prediction algorithm appears as most suitable for which task, and which type of prediction to address (e.g. how long in the future)
- **Model enhancements.** Several options are possible in this category, from exploring new data sources, to optimizing existing models for better performance, to exploring the met parameters of the models.
- **Scaling.** Start defining how for a system could be scaled by testing in different sections of the road network, and by looking at large-scale scaling.
- **Technology.** There are many technology options for NNs, ranging from training platforms, to cloud vs. on premise options, to integrating real-time data feeds, to designing GUIs and integrations.

This project addresses selected items in each of these categories.

Definitions

Definitions

- **Prediction** means estimating the status probability of road section S in time interval T with respect to the occurrence of incident of type I (incident type I does not occur or occurs ones or more times) based on data in time interval H.
    - Example: probability of a a car accident (I) in the next 30 minutes (T) on section of the road between km 10 and km 40 (S) based on data on the past hour of traffic (H).
- **Baseline** means the prediction based only on past incident data.

2

DATA REVIEW AND DATA MODELING

DATA

We use datasets provided by RWS. Specifically we use two main datasets:

- Traffic data from NDW for the period 2010-2015
- Incident data from VIAS for the period 2000-2016

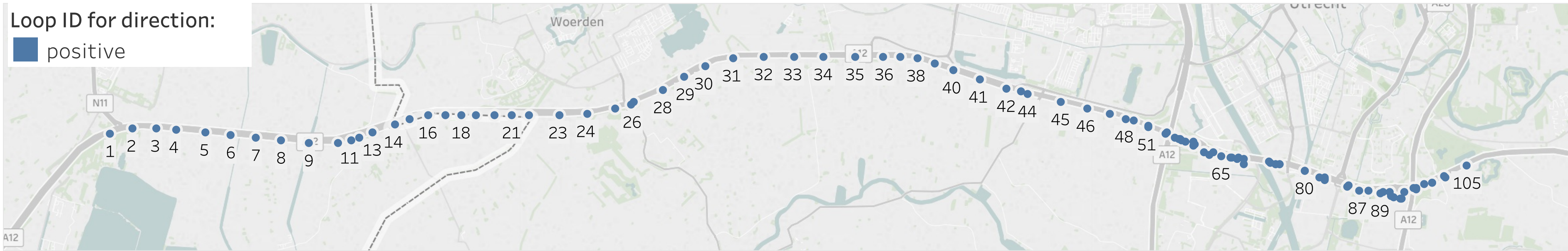
The data undergoes a process of validation, modeling, selection and processing (described in Appendix). Most data is discarded from the analysis because of incompleteness. Around 10% of the entire dataset if sufficiently complete to train a machine learning algorithm.

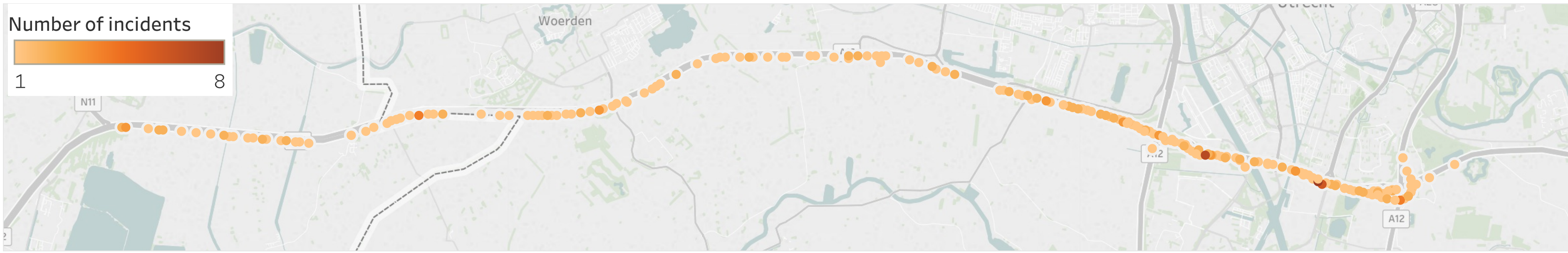
The project test area are segments of the A12 and a combined segment of A10-A4-A9-A2 south of Amsterdam.

See appendix for the processing procedure and infrastructure used to process the data.

A12 - Segment 1

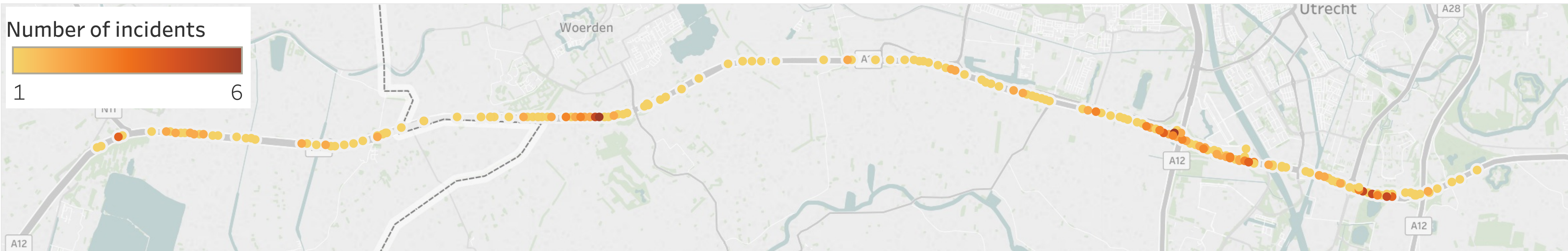
- A12 - segment 1 – loop index 1-105
- Direction – 'Re'/ positive
- Hectopoints : 34 km – 64 km
- Total number incidents in 20 months:
  - All types: 1597
  - Of which Ongeval incidents : 327
- The incidents are split into two sets, one for training and one for testing:
  - TRAIN (01.01.2013 – 30.04.2014): 236
  - TEST (01.09.2014 - 31.12.2014): 91





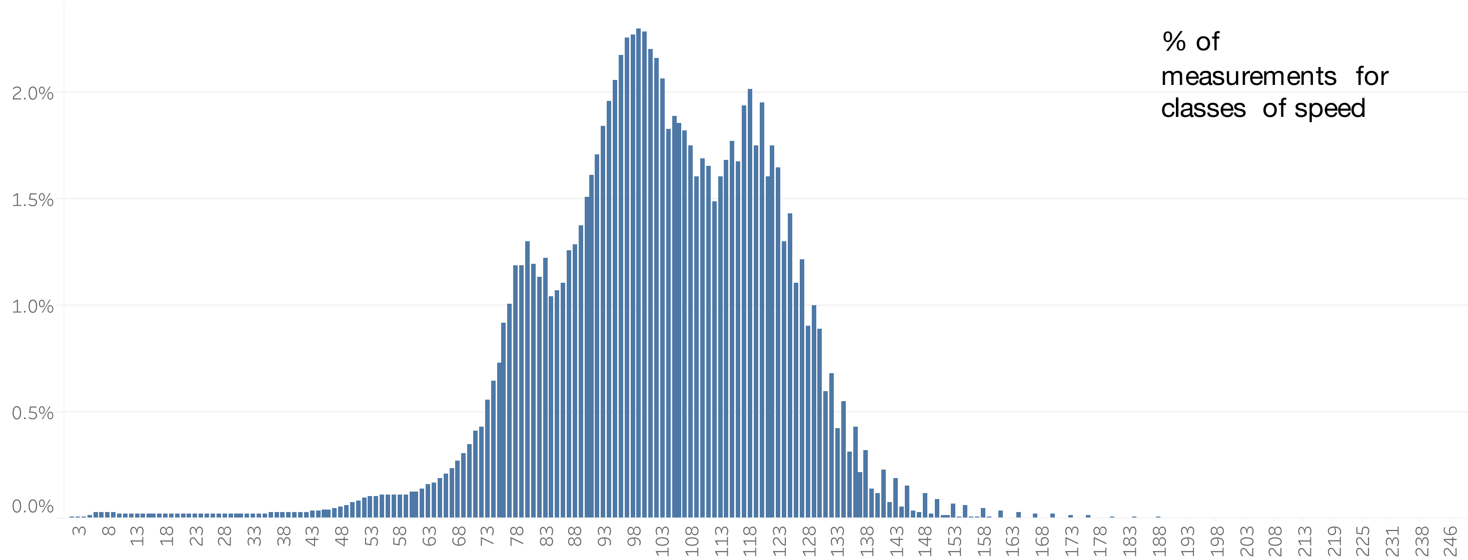
A12 – Segment 2

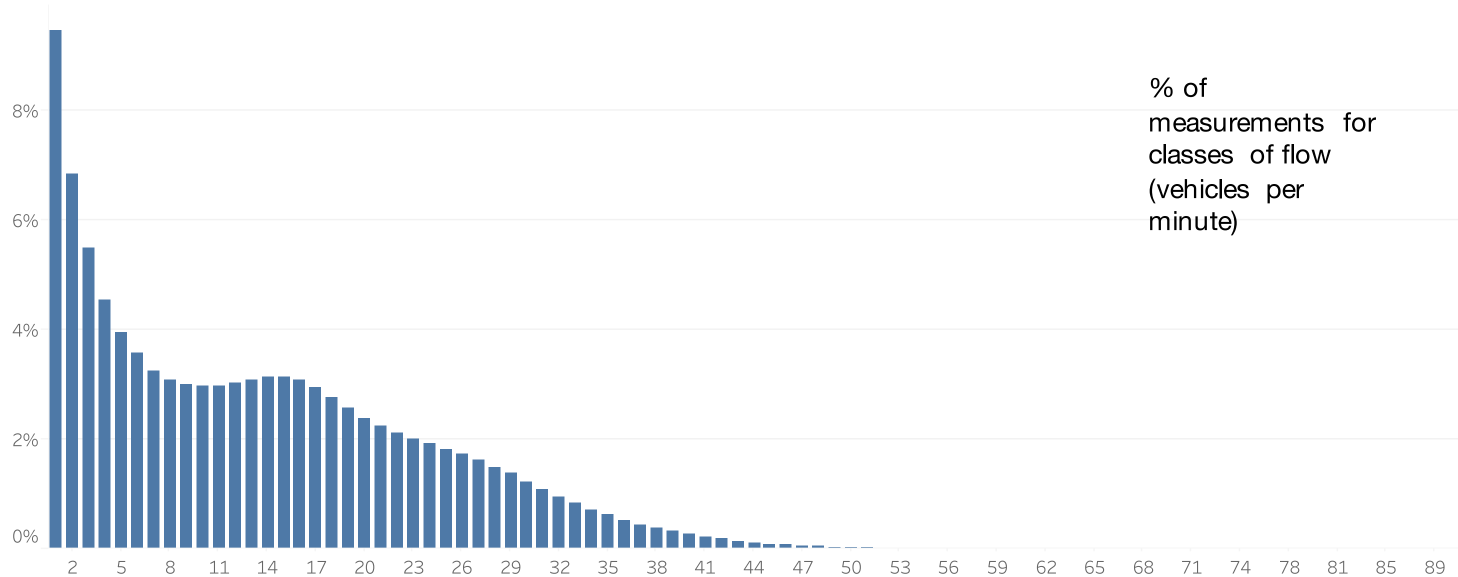
- A12 - segment 1 – loop index 99-213
- Direction – ‘LI’/negative
- Hectopoints : 34 km – 64 km
- Total number incidents in 20 months:
  - All types: 1405
  - Of which ongeval: 329
- The incidents are split into two sets, one for training and one for testing:
  - TRAIN (01.01.2013 – 30.04.2014): 250
  - TEST (01.09.2014 - 31.12.2014): 79



Speed and flow profile A12

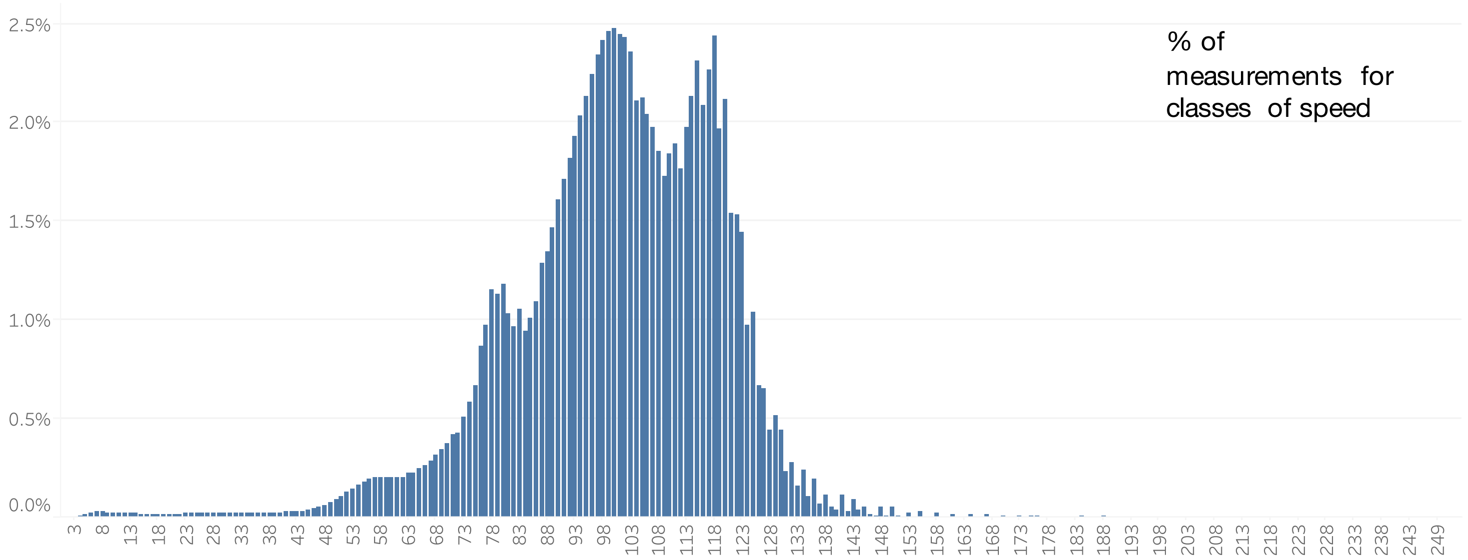
Data 2013-2014, km 34-64, Rechts



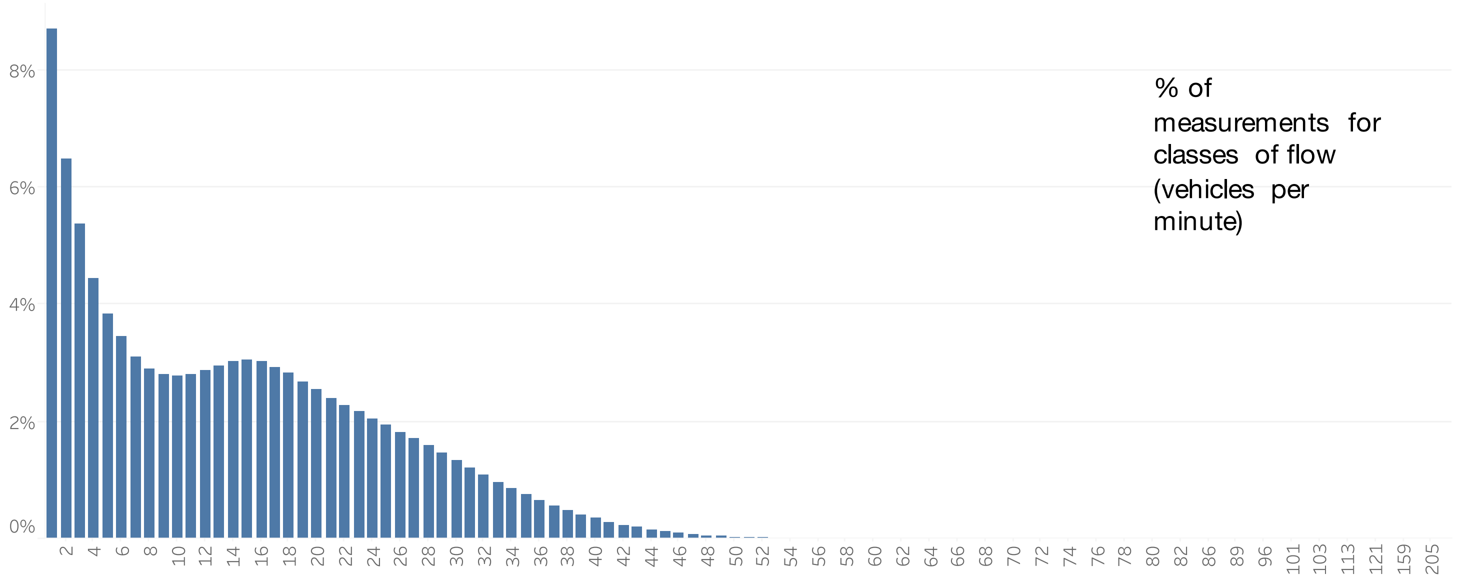


Speed and flow profile A12

Data 2013-2014, km 34-64, Links



% of measurements for classes of speed



% of measurements for classes of flow (vehicles per minute)

The general model

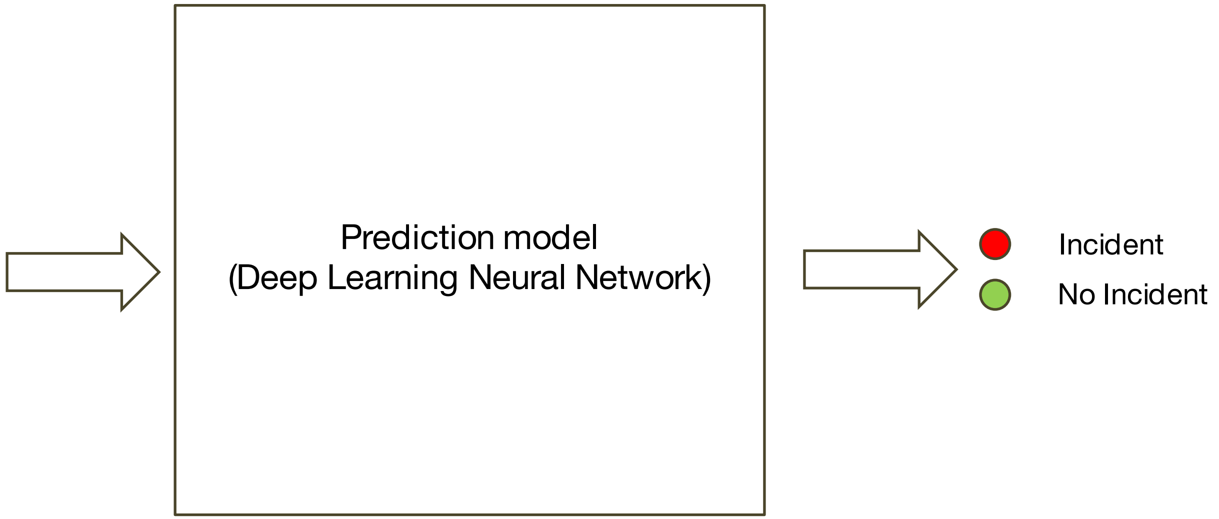
1. Traffic (and other data) are structured into images spanning a section of the highway (S) for a time block (H).

2. Images are split into training (80%) and validation (20%). The training set is labelled 1 or 0 if an incident (I) occurs (or not) in a specific time window in the future (T)

3. The NN is trained to recognize the relationship between data and labels.

4. The trained network is presented with new images to verify the prediction capability.

	T0	T0+ 10m	T0+ 20m	T0+ 30m	T0+ 40m	T0+ 50m
L1	Speed Flow					
L2	…	…	…	…	…	…
L3						
L4		Speed Flow	Speed Flow	Speed Flow	Speed Flow	
…						
….						
Ln						Speed Flow



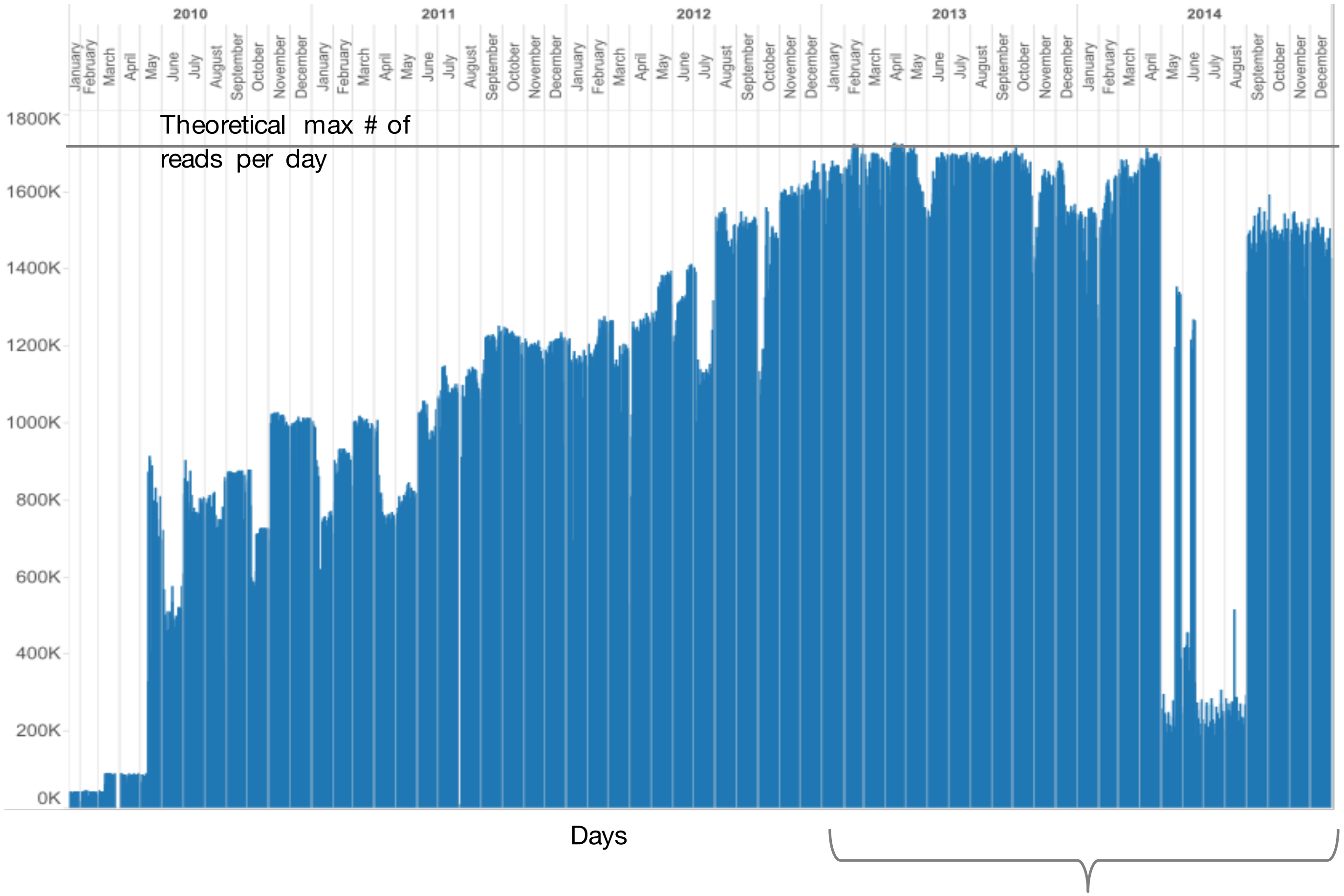
Traffic data: completeness of traffic measurements in time

The diagram shows the number of ”good” reads for traffic loop data per day in the period 01/2010 – 12/2014.

The original data format is
- speed-flow
- per loop-lane
- per minute

In this dataset, assuming all loops generate good data continuously (every minute), we would have 1,715,040 maximum reads per day.

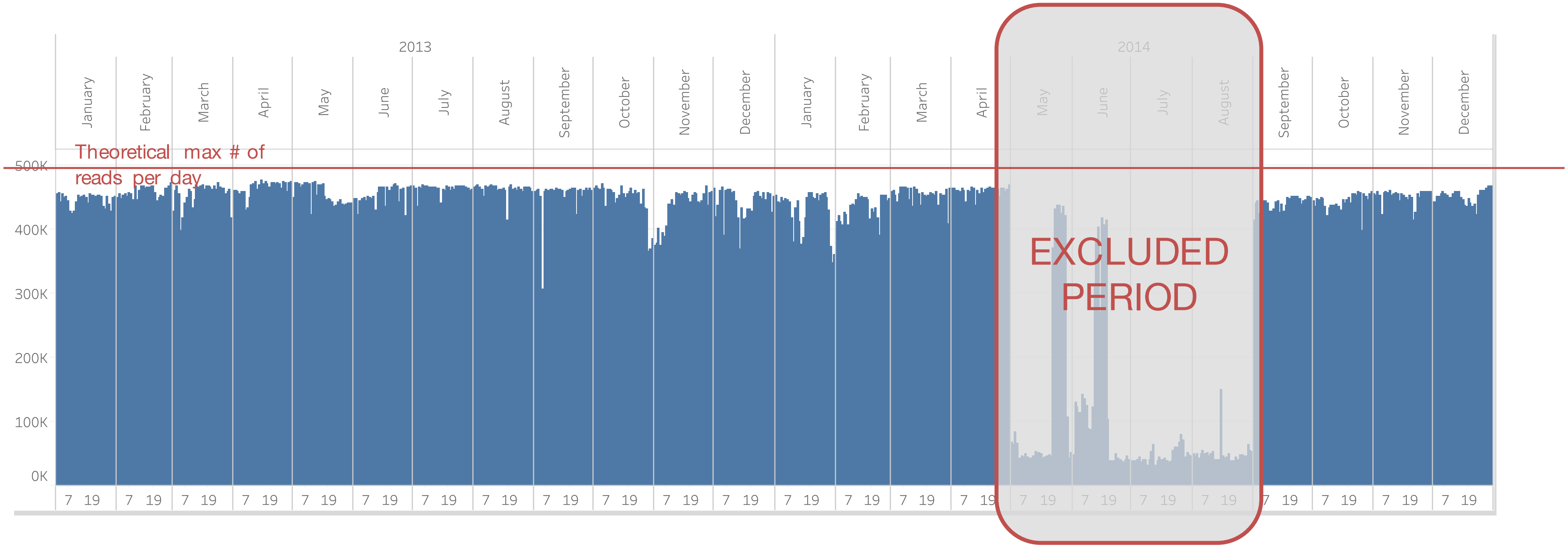
The average of good reads over 5 years in this dataset is 1,140,465. The % good reads per day is 66.5%.



In this project we use the most compete portion of the data between January 2014 and December 2015, excluding the May-August gap in 2014.

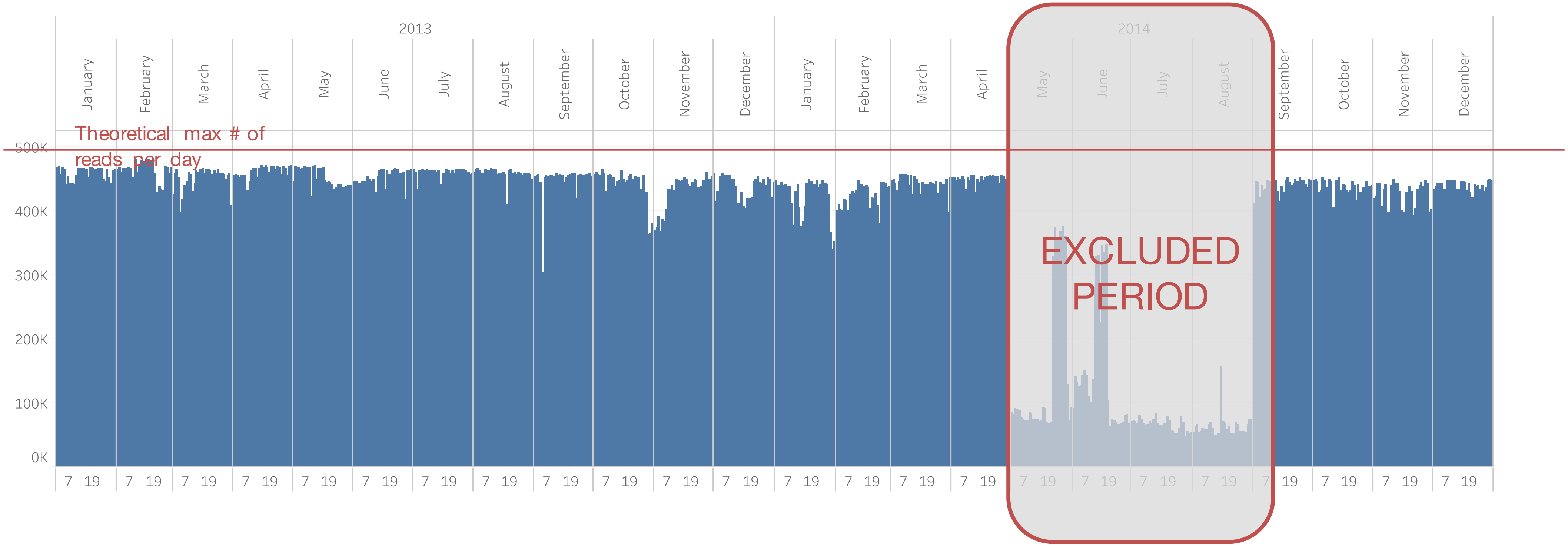
Traffic data: A12 segment 1

The diagram shows the number of ”good” reads for traffic loop-lane data per day in the period 01/2013 – 12/2014 for selected segment on the A12. For this segment, assuming all loop-lanes generate good data continuously, we would have maximum **521,280** reads per day. The average good reads per day for this period is **450,476** data points (**86 %**).



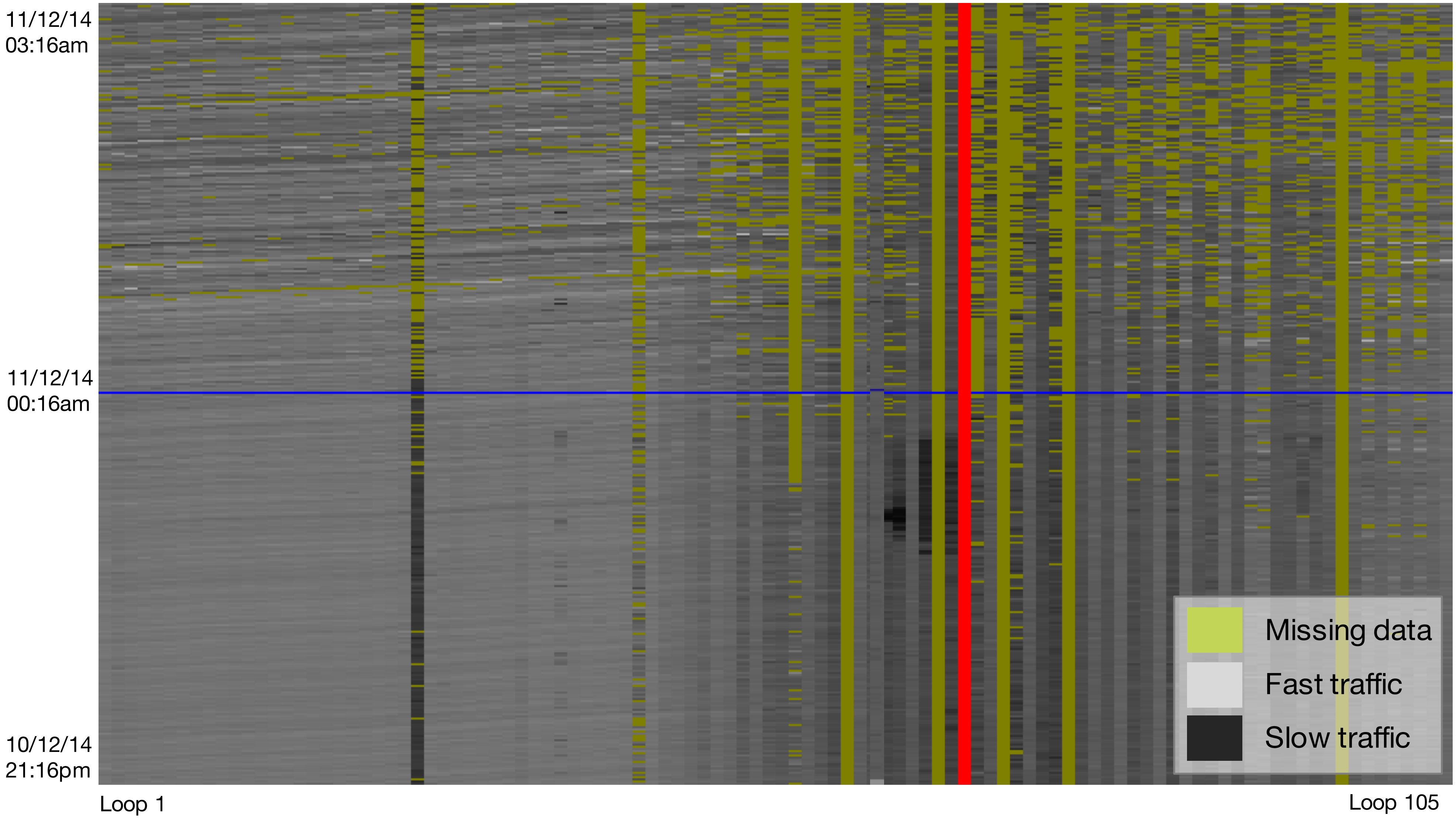
Traffic data: A12 segment 2

The diagram shows the number of ”good” reads for traffic loop-lane data per day in the period 01/2013 – 12/2014 for selected segment on the A12. For this segment, assuming all loop-lanes generate good data continuously, we would have maximum **537,120** reads per day. The average good reads per day for this period is **444,719** data points (**83 %**).



Traffic data: example of time-space completeness

A12, km 34-64, Rechts



Analysis of traffic data

The system uses images constructed from traffic reads at loops at different times. For the period of 20 months used in the project:

- There are potentially 864,000 images of traffic speed and of traffic flow (one per minute), each spanning a fix time window (e.g. 30 minutes)
- From the available dataset **none** of these images are complete (contain data for all loops for the entire time window)
- For A12
  - Re: ± 457,000 images are least 90% complete (±52% of the available images)
  - Li: none are 90% complete, ±600,000 images are at least 80% complete

- Implications:
  - Since the prediction algorithm scans for patterns from sequential images, ideally each incident is associated to a continuous set of images. In reality, none of the incidents has a continuous set of images for training.

Incident data

We have access to incident data from 2010 to 2015 from VIAS.

The dataset records a variety of data points for each incident. The variables that are essential for training a prediction model are:

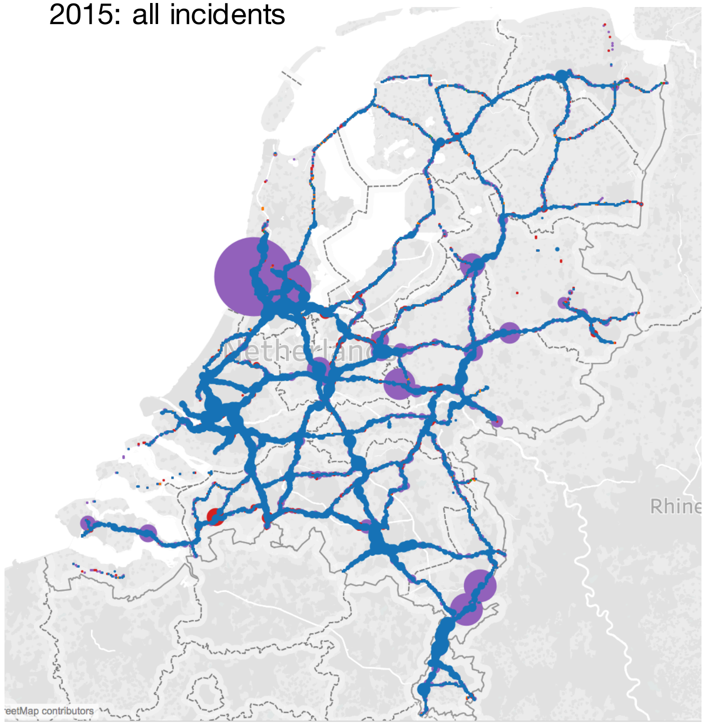
- The incident classification (ongeval, gestrande voertuig, …)
- The incident start time (when the event has taken place)
- The incident geographic location (start and end of area affected)

Additional attributes that are useful for incident interpretation are:

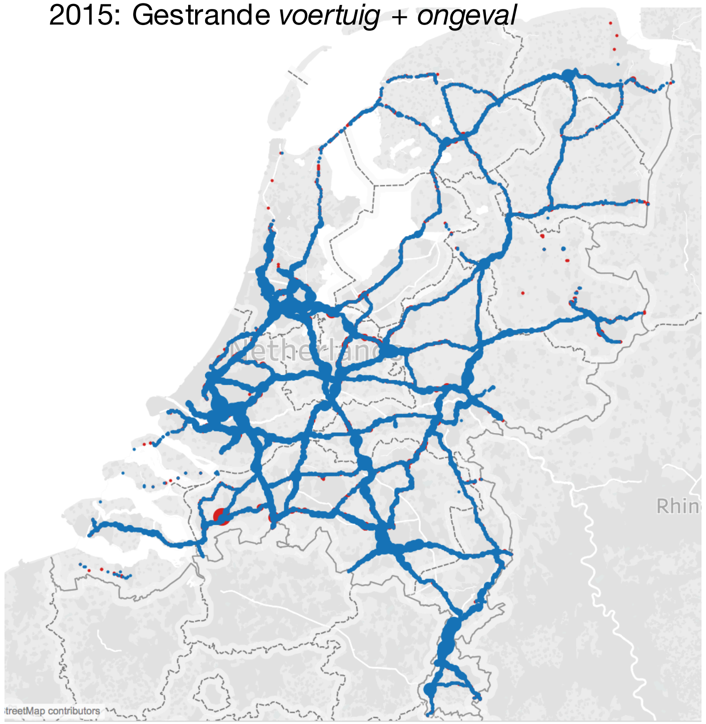
- Incident duration
- Incident description (a proxy of the extent of the incident consequences)

Incident data: overview example

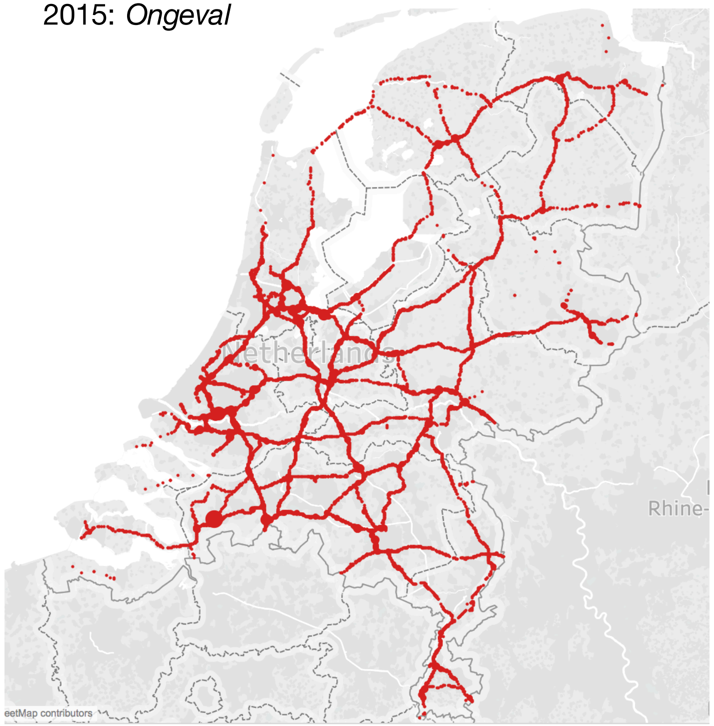
2015: all incidents

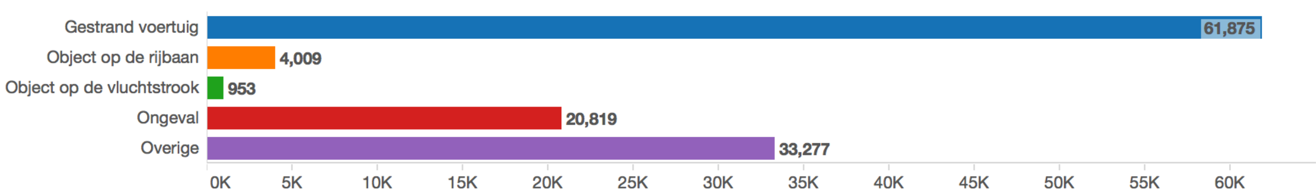


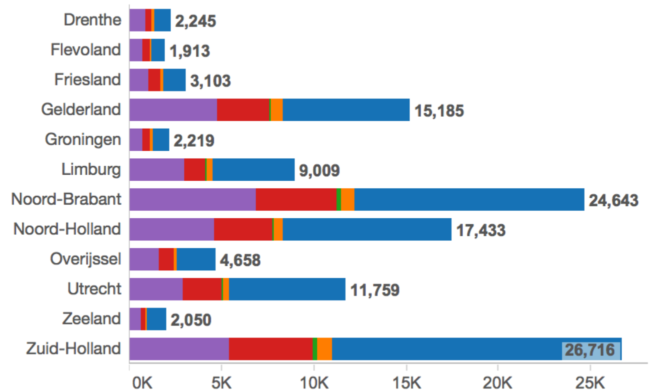
2015: Gestrande *voertuig* + *ongeval*



2015: *Ongeval*

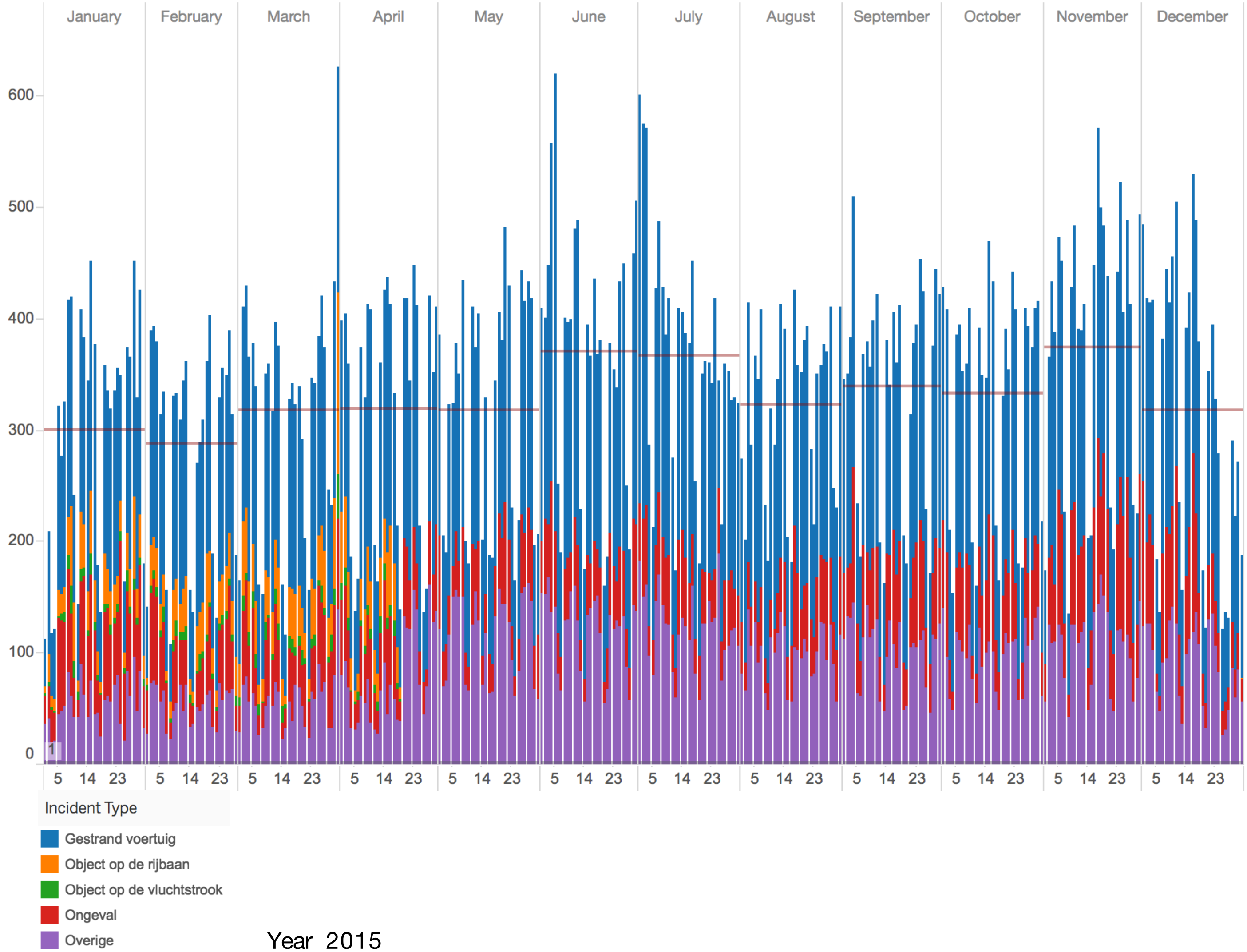






Year 2015

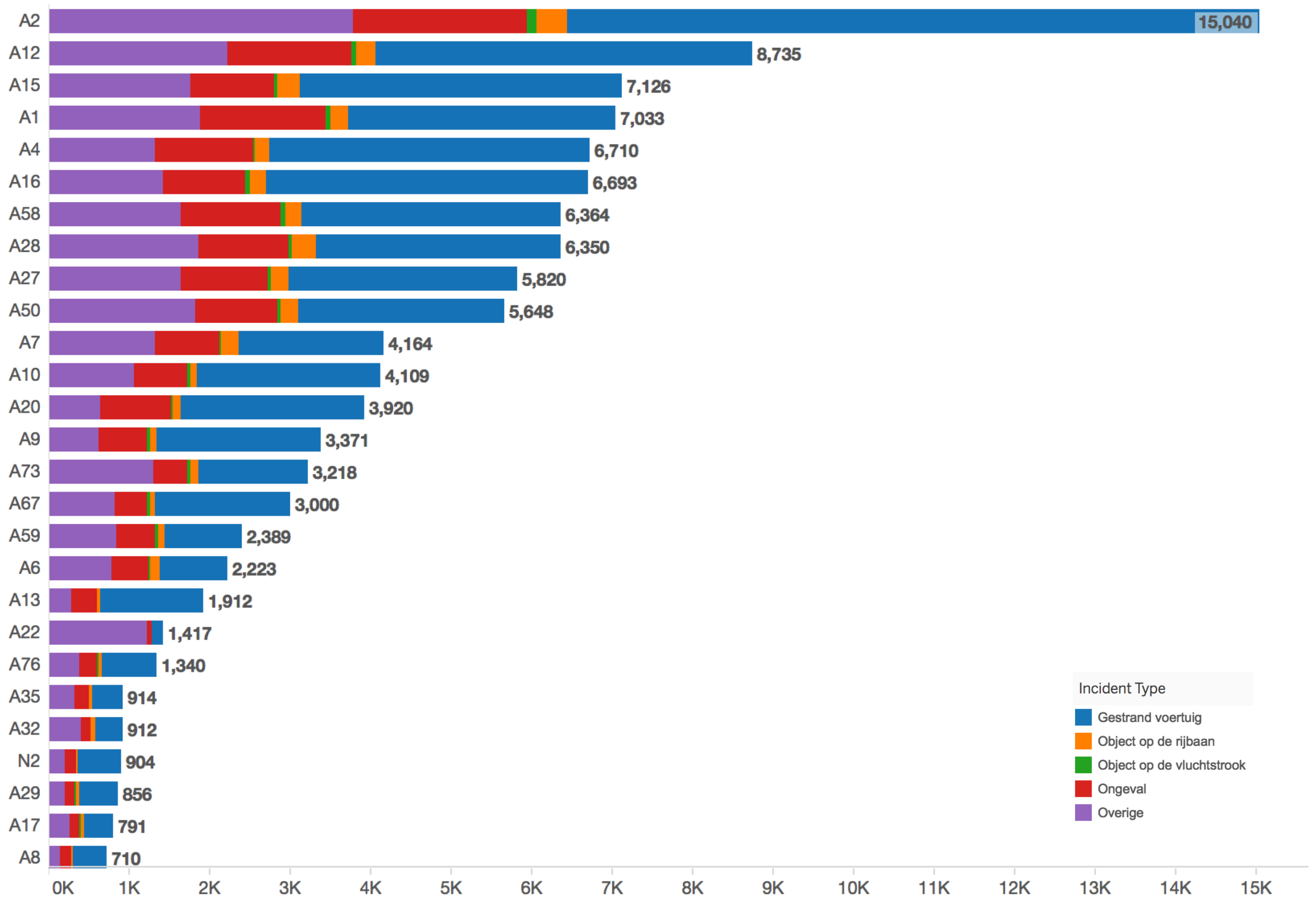
Incident data: overview (cont.)



Year 2015

bestelbus	292
Meerdere	2,073
motor	138
Onbekend	113,999
personenauto	3,542
tankwagen	39
touringcar	31
vrachtwagen	819

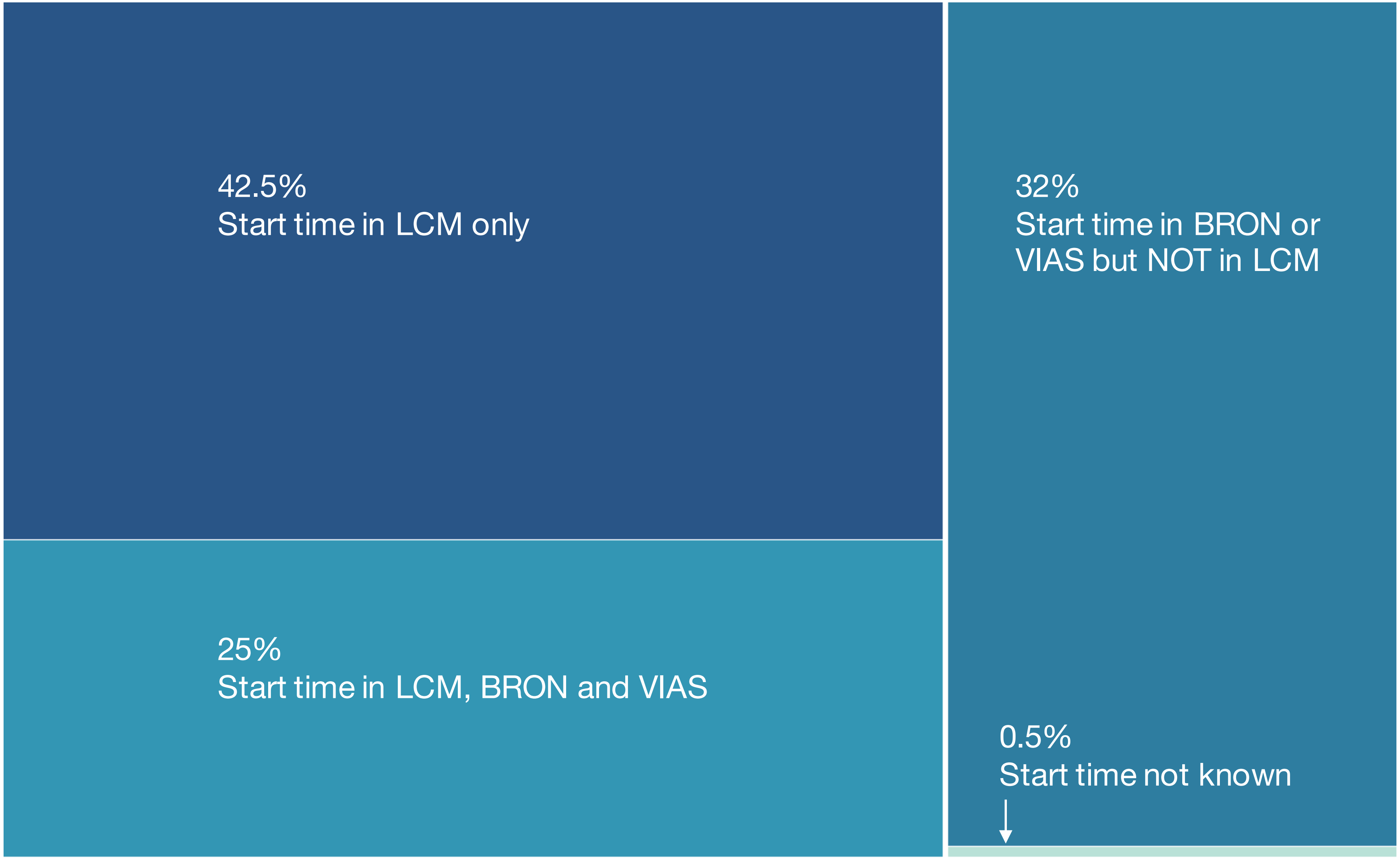
Incident data: overview (cont.)



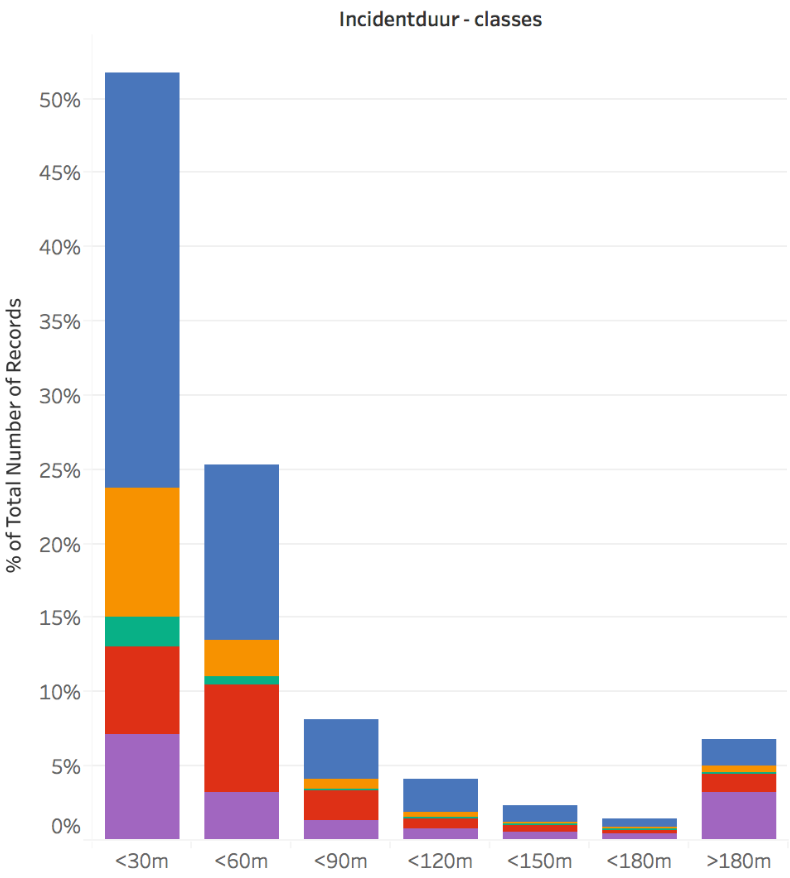
Year 2015

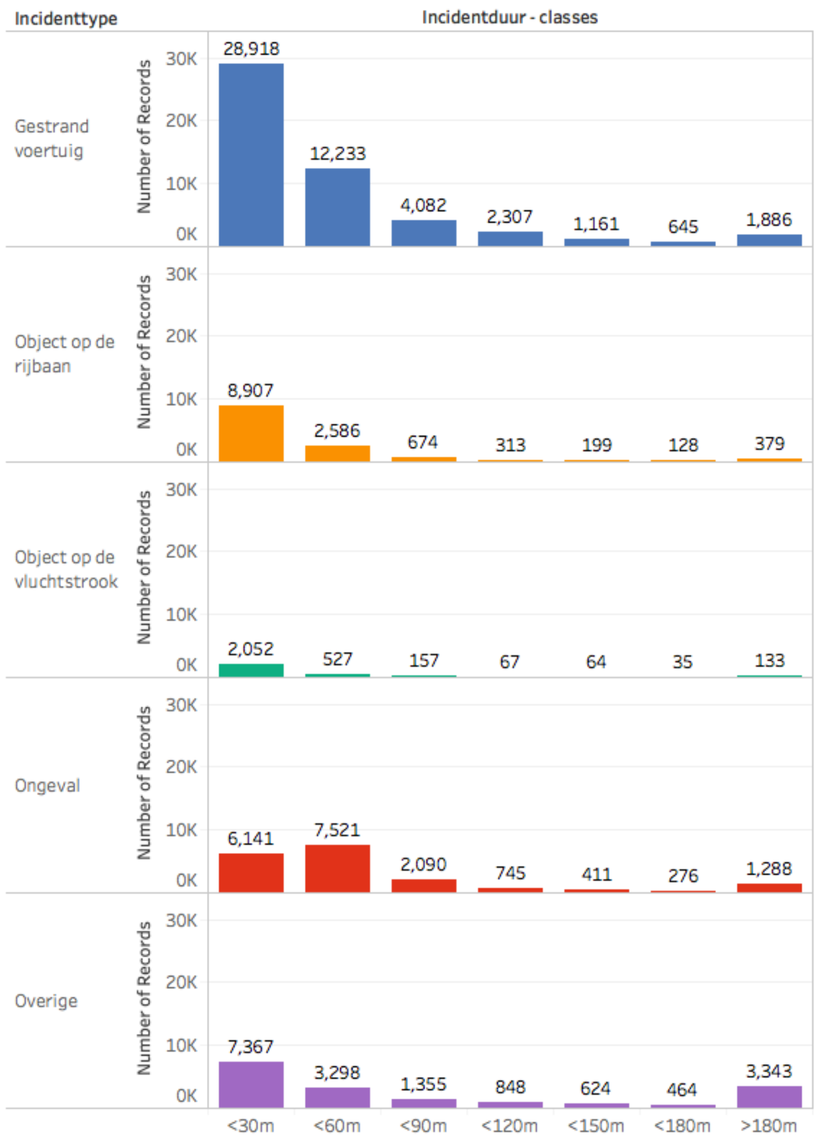
Incident start time: accuracy

Period 2010-2015
Sources: BRON, CMV, LCM (Bergers), UDLS (VIAS)
Total # incidents: 263,557



Incident duration in classes





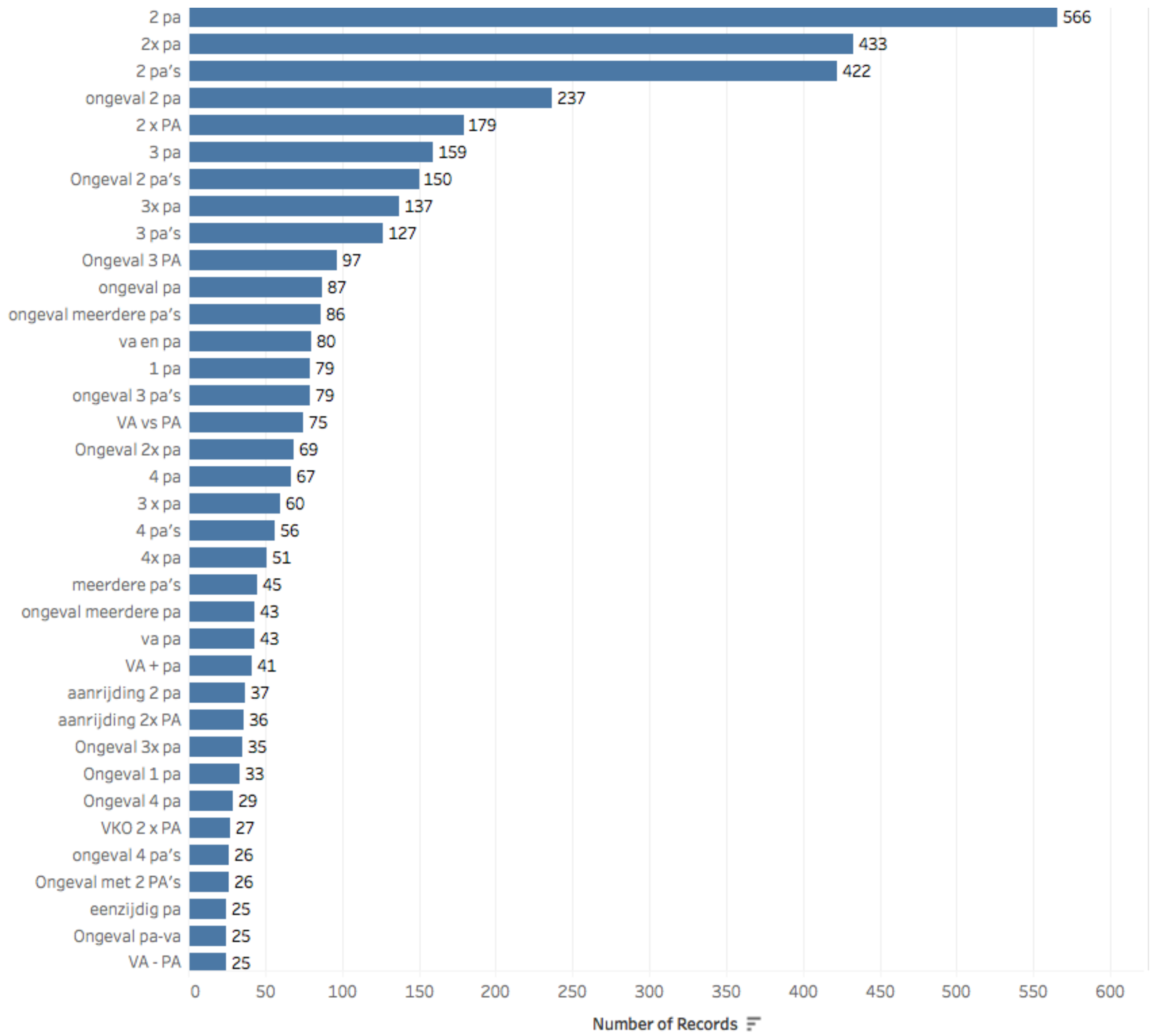
Year 2014

Incident description

Year 2014:

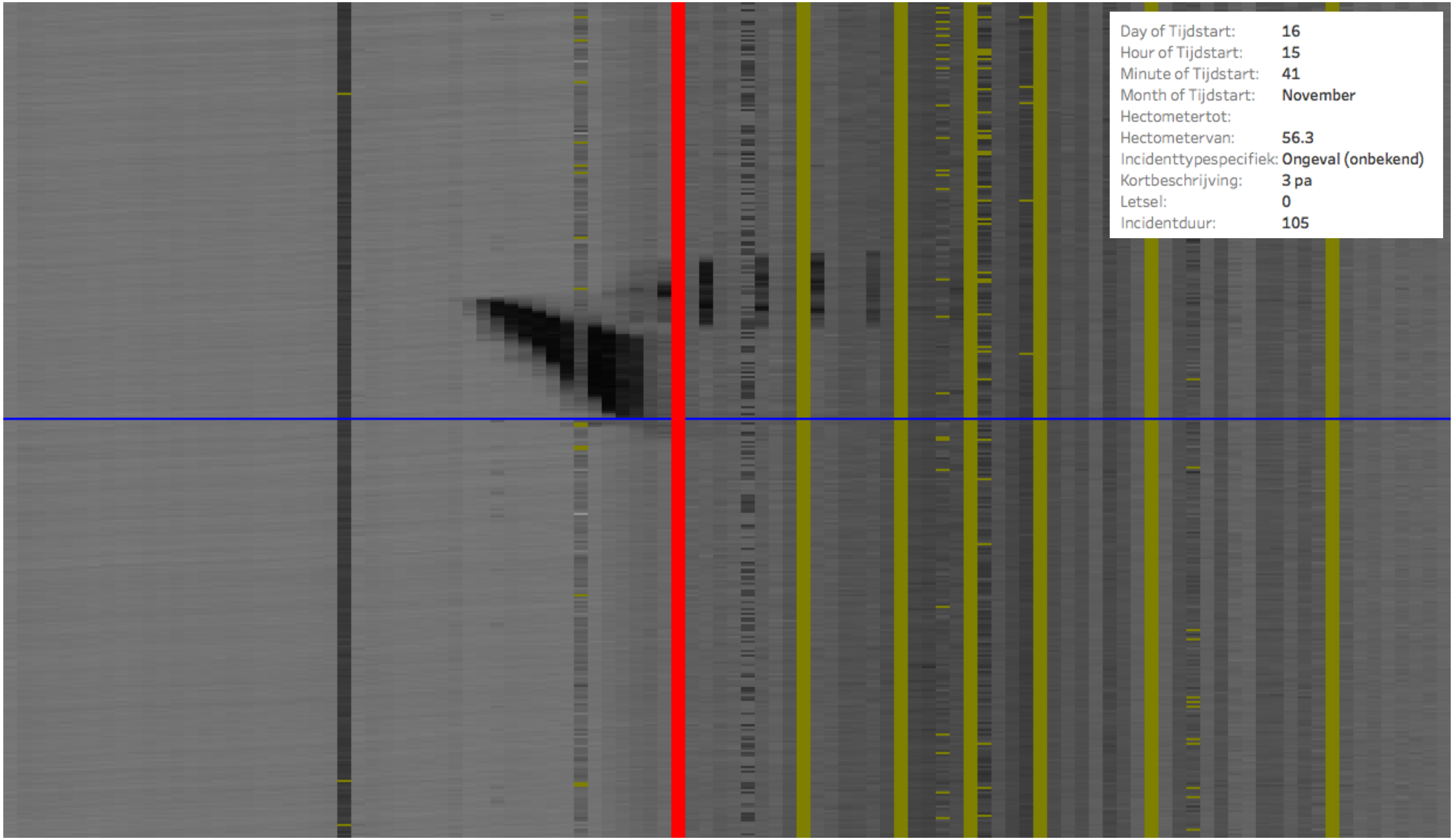
- 18,472 ongevallen
- 1,042 events do not have a description
- 1,192 have the general "ongeval" description

The rest have variety of inconsistent descriptions. For instance, the term "PA" appears in 6,927 records in 1,778 variations.



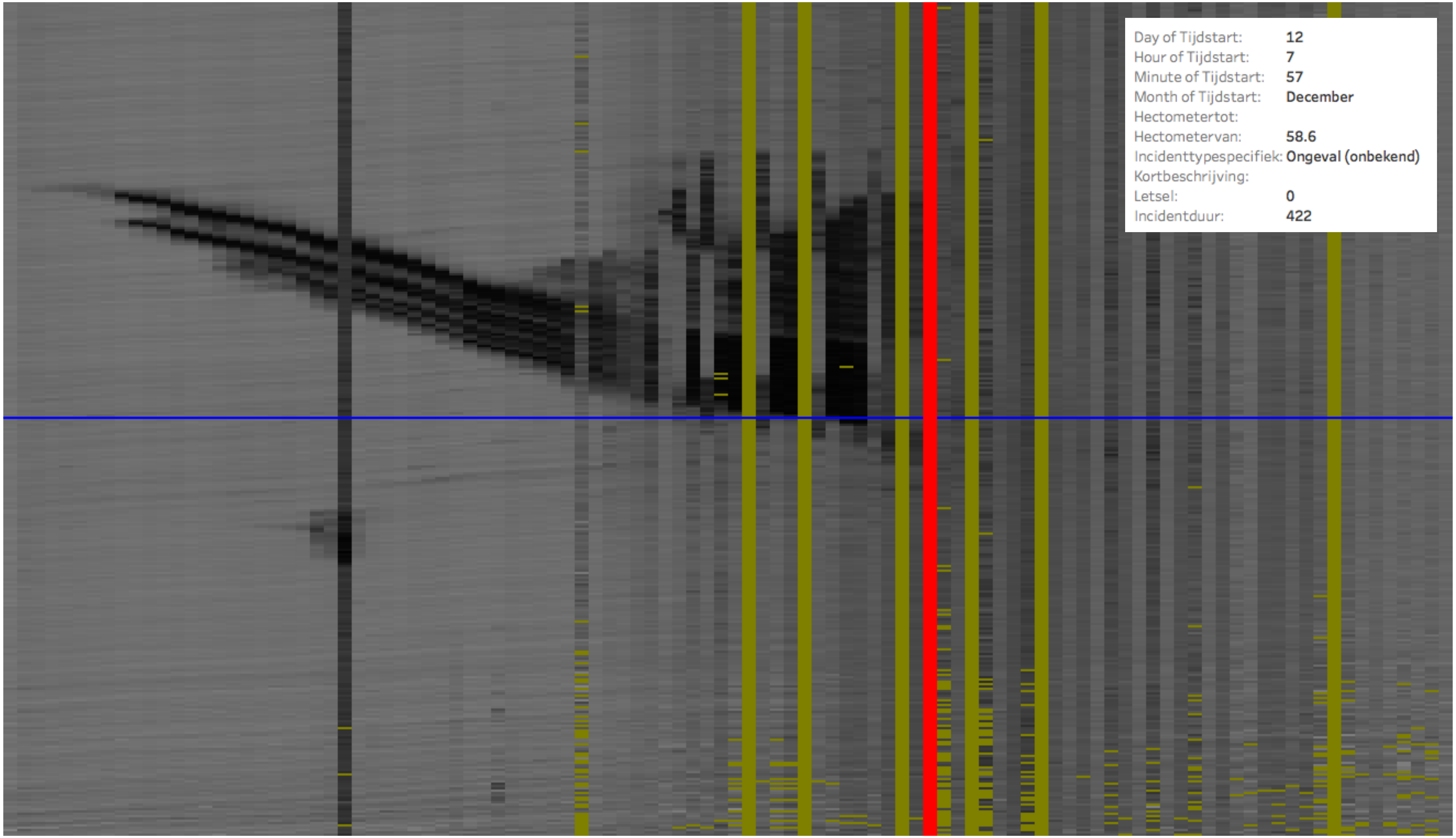
Incident start time validation: probably right

Starts at the time and place of incident. The disturbance is compatible with an incident duration of 1 hour.



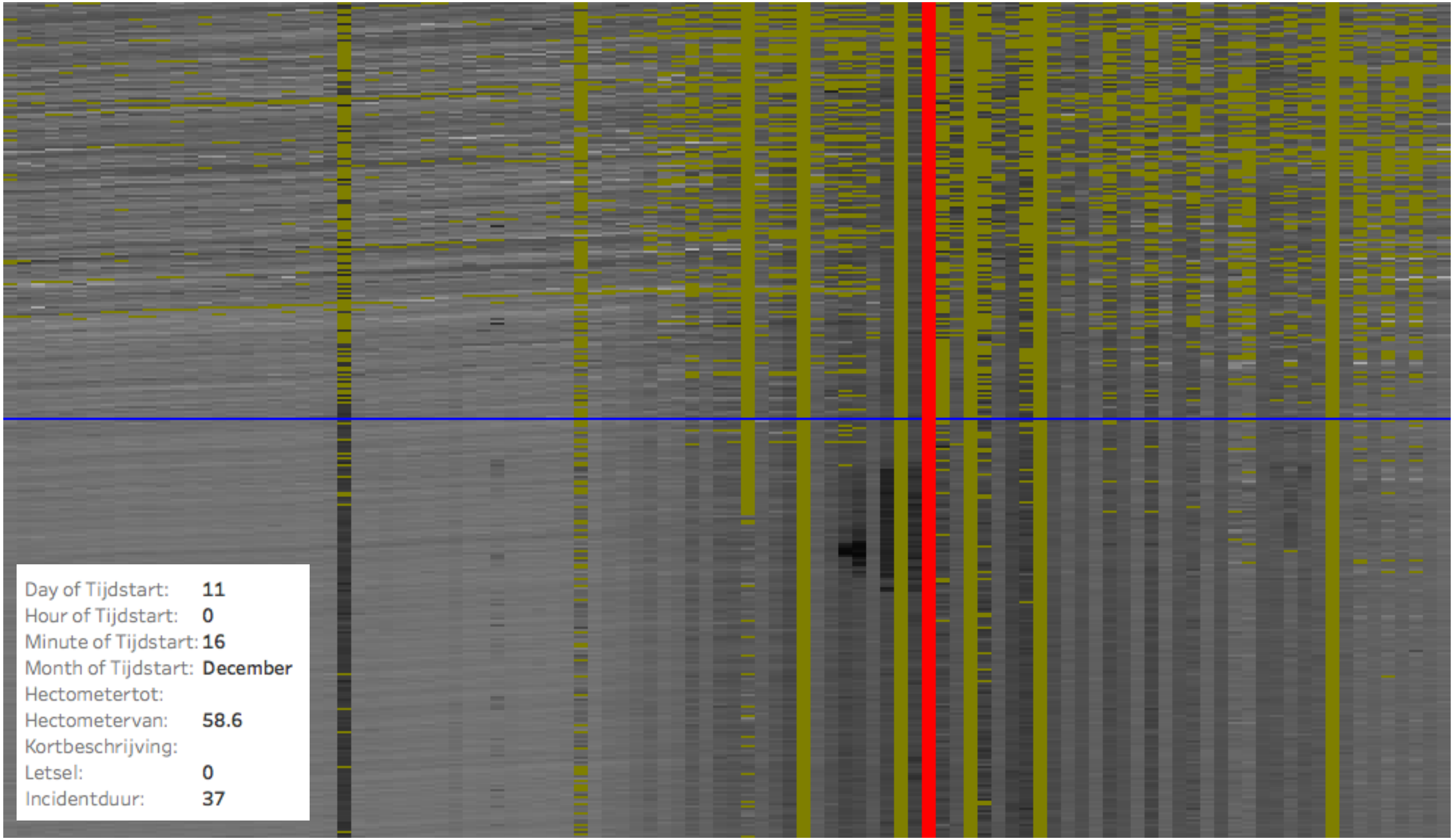
Incident start time validation: correct?

Traffic jam starts before incident. One reason could be time registration of incident later than incident time.

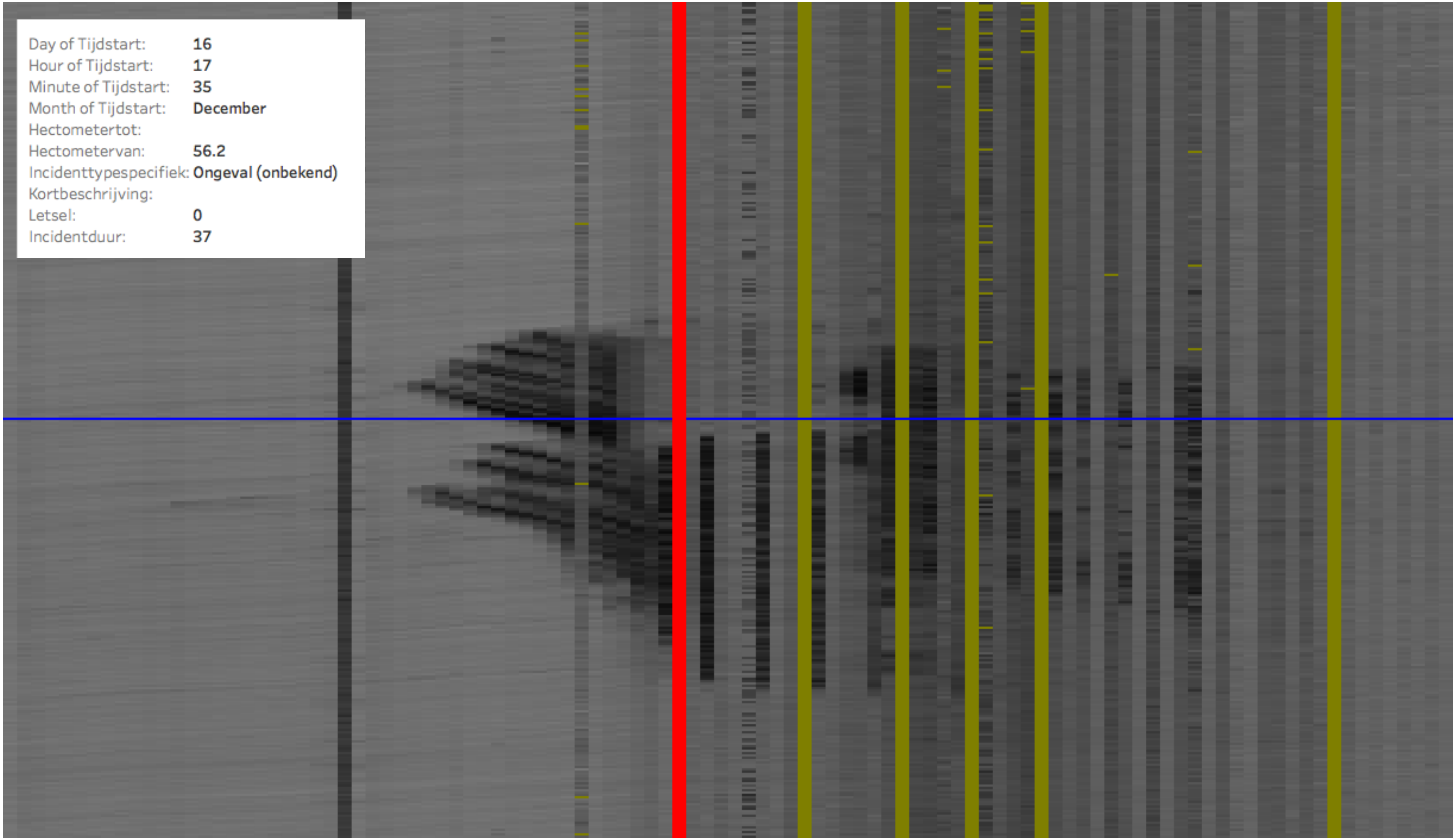


Incident start time validation: correct?

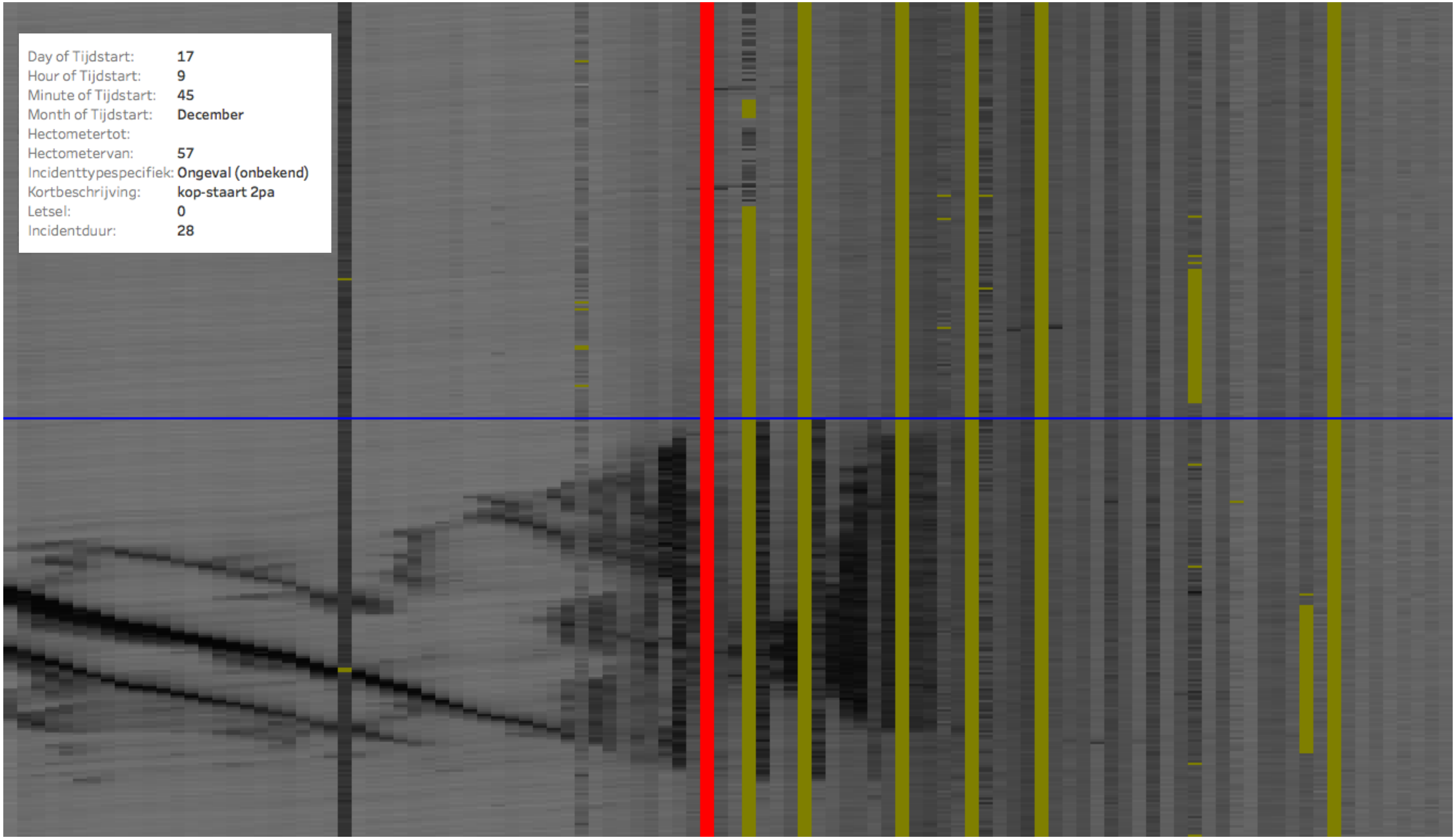
Disturbance before incident but no disturbance at incident (is it plausible that the incident did not cause a noticeable disturbance?)



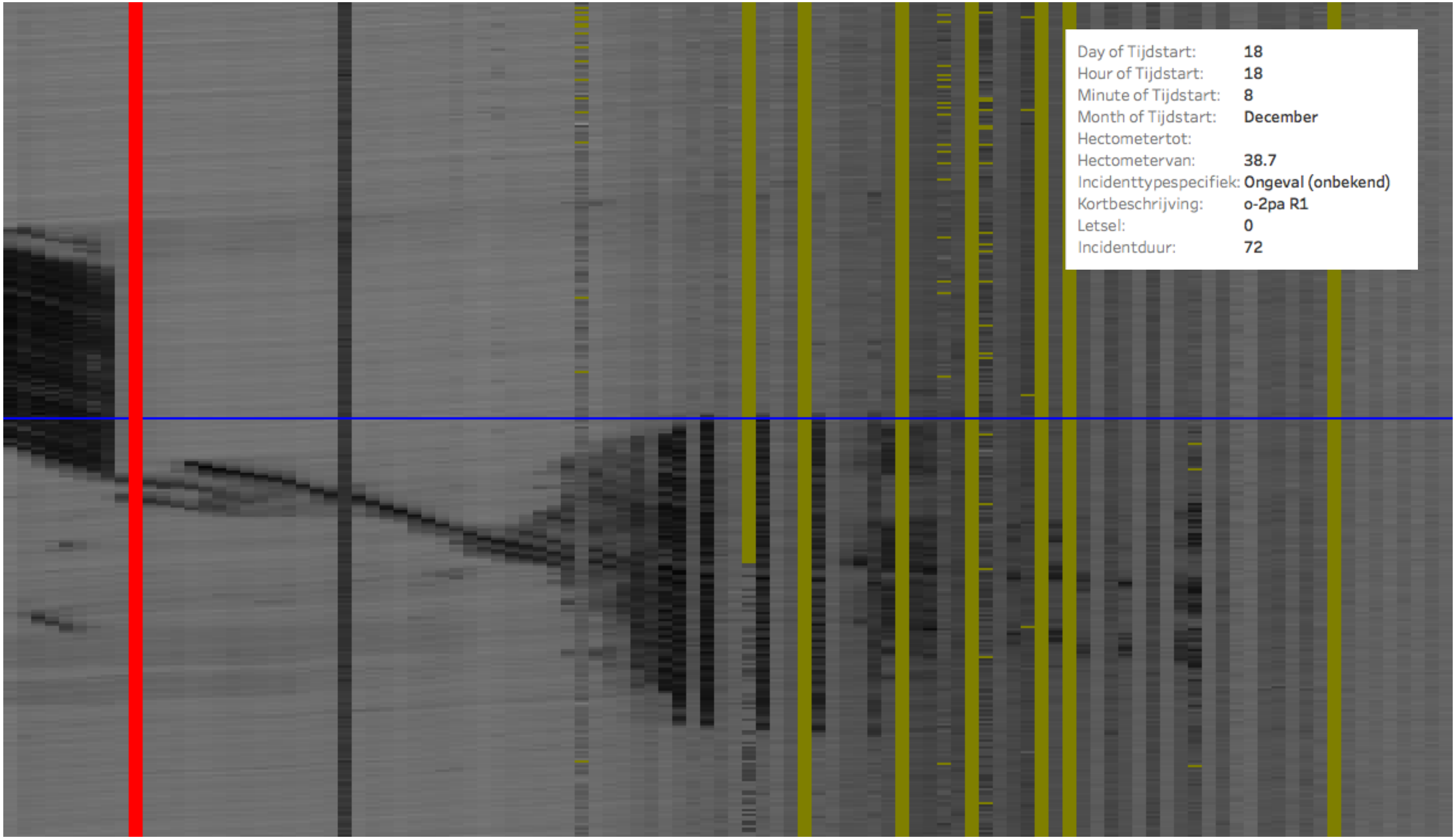
Incident start time validation: correct?



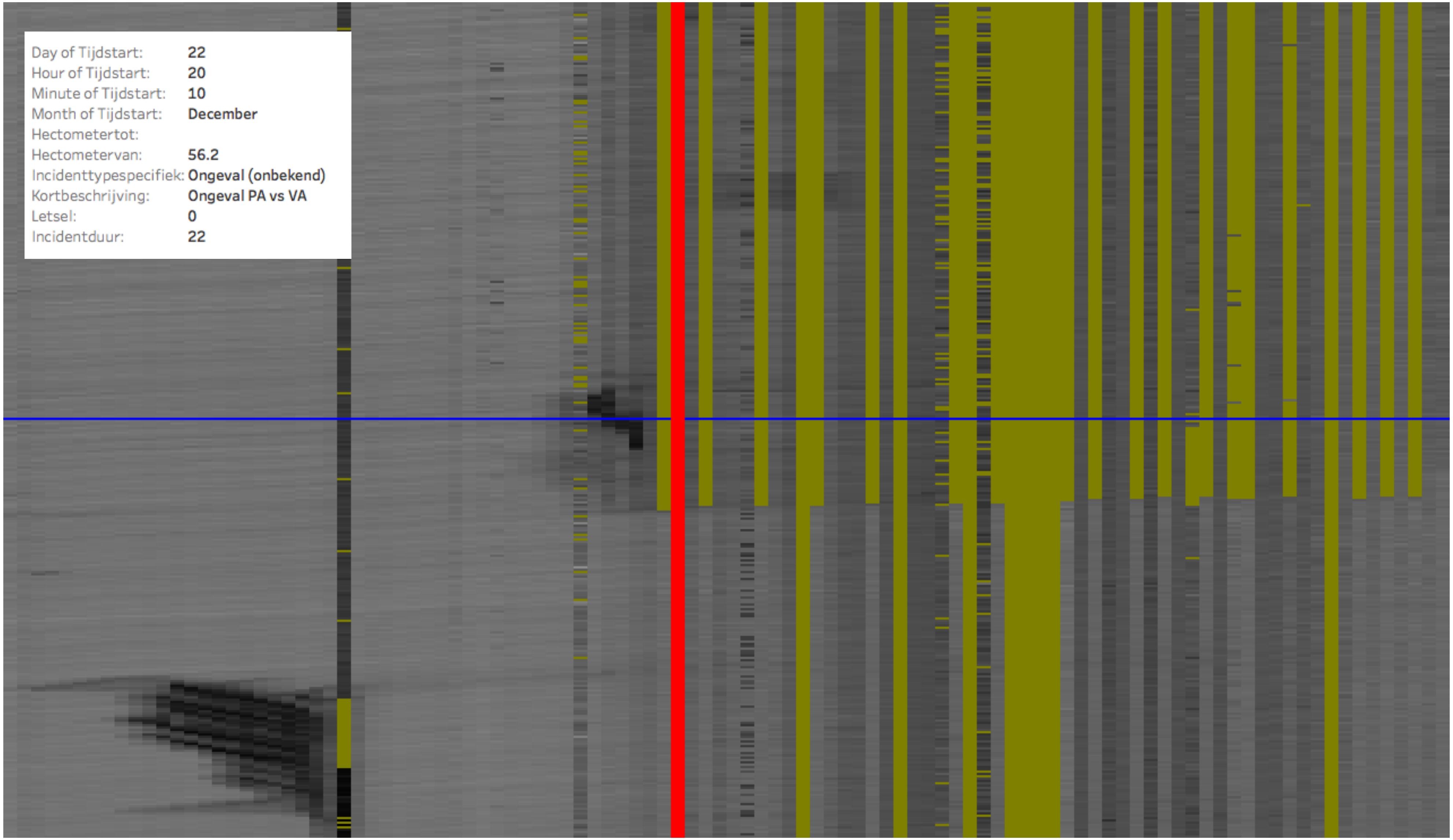
Incident start time validation: correct?



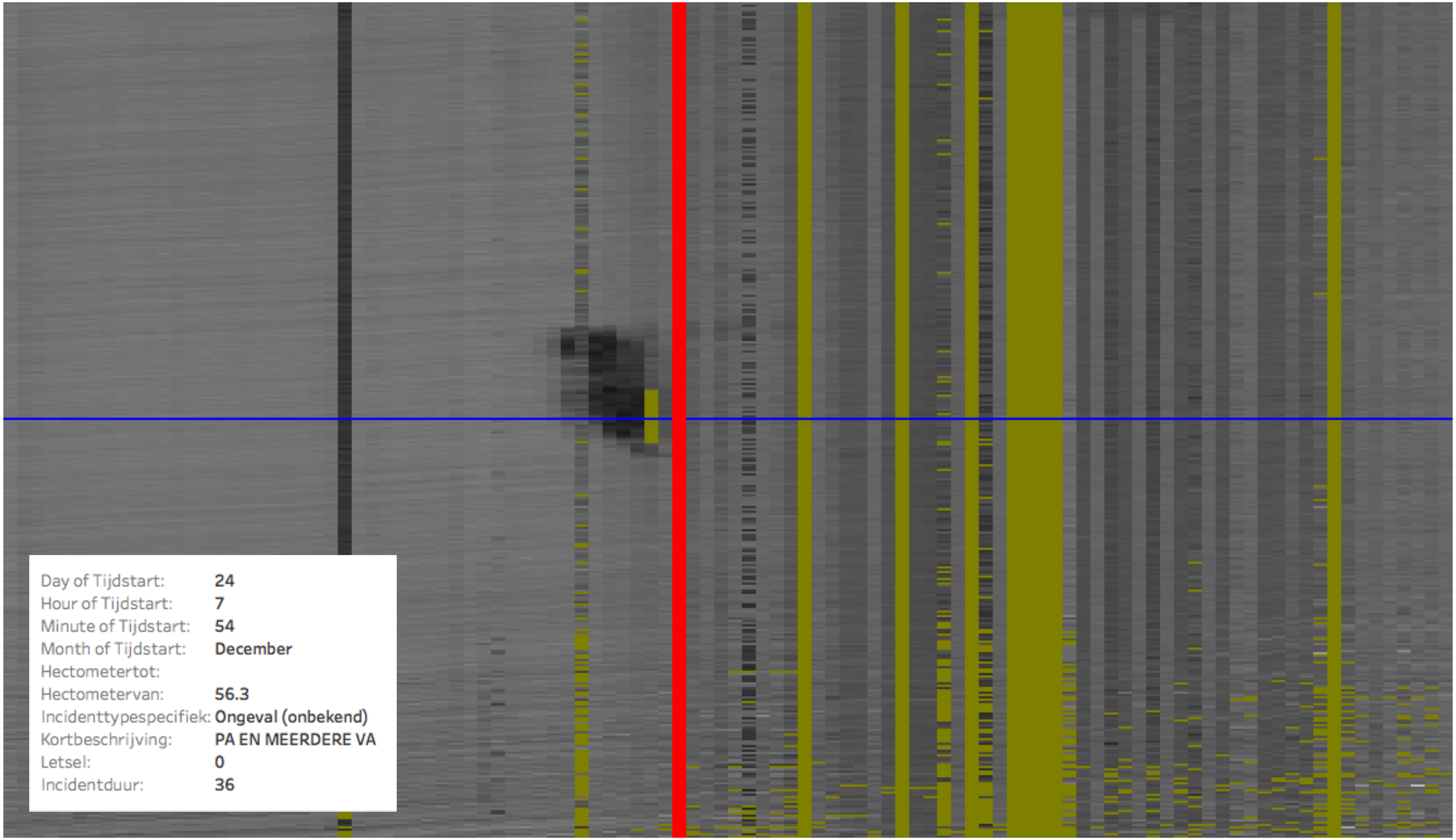
Incident start time validation: correct?



Incident start time validation: correct?



Incident start time validation: correct?



Analysis of incident data

- There are 4 main questions on the incident dataset:
  - Does the list of incidents contain the full list of actual incidents that took place? If the list is incomplete, then the machine learns patterns that are associated to a real incident but are not considered as such because of missing data.
  - Is the start time of the incident correct? If the start time is incorrect, the machine learns patterns that may or may not be precursors of an actual incident, or possibly learn patterns that occur after an incident.
  - Is the location of the incident correct? Imprecision in location registration is not immediately crucial for the prediction model (the model predicts over a stretch of the road, not on a specific point) but it is essential for data interpretation and forensics.
  - Does the incident duration reflect the duration of the traffic disturbance? Imprecision in duration registration is not immediately crucial for the prediction model (the model predicts the start time window) but it is essential for data interpretation and forensics.
- The incident start time is a major issue: in many cases it appears as approximate.
- The incident duration has a vast spread with a significant # of incidents with a very long duration (> 180m): it is unclear if duration is precisely measured and if duration is interpreted in the same way across all incidents.
- The incident specific description (which indicates the severity of the incident in qualitative manner) is a free field with a large number of description style. To be used it requires a major cleaning effort.

Observations on implications of data completeness

Traffic data incompleteness impacts the learning of NNs in multiple ways:

- It reduces the data pool that can be used to train the network
- It reduces the ability of detecting patterns

Incident data issues affect both the models and the interpretation of results:

- Errors in incident time registration affect the labeling of the training sets, potentially feeding the algorithm with wrong incident labels (this triggers the NN to seek for wrong patterns).
- Incomplete or incorrect incident detail data makes it very difficult to reconstruct the precise incident conditions and to interpret predictions results.

There are two main strategies for addressing data incompleteness and inconsistencies:

- At the model level
  - Introducing robustness (e.g. through time buffers) to reduce impact of time incident registration
- At the data level
  - Data exclusion: remove all images that are less than 90% complete
  - Lateral aggregation (across lanes): removes data gaps (e.g. single lane off)
  - Longitudinal aggregation (across loop locations): removes data granularity and gaps
  - Linear data interpolation (to fill in missing traffic data): applicable to small gaps
- Long term solutions:
  - Local smoothing: apply smoothing rules to neighoborhood sets of data with methiods from image processing. This may require multiple differnet rules for different areas
  - LSTM interpolation (use of a sequential NN to learn to fill in the actual dynamics of the road)

3

DEEP LEARNING NEURAL NETWORK MODELS AND RESULTS

Models and their design parameters

Training data	Traffic		Traffic and weather	
Time window in the future (T)	15 m	30 m	60 m	120 m
Rolling history (H)	15 m	30 m	60 m	120 m
Road section length (L)	30 km		50 km	
Incident type (I)	Ongeval	Ongeval + Gestrande v.	All incidents	
Traffic data granularity	1 m full resolution	1 m lateral aggregation	10 m lateral aggregation	
Highway area	Main carriage		Carriage plus exists	
Neural Network design	FC - Fully connected	CONV 2 layers	CONV 3 layers	
Output tuning	Balanced	Avoid false positives	Avoid false negatives	

The table illustrates the main model parameters tested.

Notes:

- T is the time window in the future for which predictions are made. T=30m, for instance, means predicting the status of the road (incident or no incident) in the next 30 minutes
- H is the rolling history, that is the duration of observations (from 15 minutes to 120 minutes) which are used to predict the status of the road in the future.
- Traffic granularity indicates the level of aggregation of data to overcome gaps and incompleteness
- Neural network design is mainly Fully Connected (FC) and Convolutional (CNV)
- The output tuning indicates if the model is tuned to predict as good as possible or to privilege for instance avoidance of false positives.

Classes of models tested

Model class	**M0** FC NN. Applied to all incidents and to long T and H	**M0+** CNV with same data setup of M0 applied	**M1** CNV applied to various H and T and incident times	**M2** CNV applied to short H and T and various data resolution types	**M3** FC NN applied to short H and T for *ongeval* and aggregated data
Training data	Traffic and weather data	Traffic data	Traffic data	Traffic data	Traffic data
Time window in the future (T)	1hour, 2 hours	1 hour, 2 hours	15m - 2 hour	30m-60m	30m
Rolling history (H)	1hour, 2 hours	1 hour, 2 hours	15m - 2 hour	30m-60m	30m
Road section length	48km, 24km	48km, 24km	48km, 24km	30km	30km
Incident type (I)	All incidents	All incidents	All incidents	Ongeval, Gestrande	Ongeval
Data granularity	Lateral lane aggregation	Lateral lane aggregation	Lateral lane aggregation	Full resolution and Lateral lane aggregation	Lateral lane aggregation
Highway area	Main carriage	Main carriage	Main carriage	Main carriage and exits	Main carriage and exits
Neural network design	Fully connected	Convolutional	Convolutional	Convolutional	Fully Connected
Output tuning	Balanced	Balanced	Balanced	Balanced and false positives	Balanced

We tested models based on a variety of parameters, broadly falling into 5 classes. In the process we repeated tests made in previous projects of RWS on the same data.

The project has deployed more than 50 individual models with various combinations of design parameters. 24 of these models are reported on next page, as representative of the full set of tests.

Summary of main models tested

(Selected set from the full set tested)

Model	NN type	Incident type	History (months)	Segment length (km)	Prediction Time window (m)	Rolling history (Image time window, m) - H	Data granularity	exit-no exit	Time resolution traffic data (m)	Buffer incident time (m)
M0-1	FC	all	20	48	120	120	Lateral aggregation	no exits	10	0
M0-2	FC	all	20	24	60	120	Lateral aggregation	no exits	10	0
M0-3	FC	all	20	12	60	240	Lateral aggregation	no exits	10	0
M0-4	FC	all	20	6	60	240	Lateral aggregation	no exits	10	0
M0+ -1	CNV time-space	all	20	48	120	120	Lateral aggregation	no exits	10	0
M0+ -2	CNV time-space	all	20	48	120	120	Lateral aggregation	no exits	10	0
M1-1	CNV time-space	all	20	48	120	120	Lateral aggregation	no exits	1	0
M1-2	CNV time-space	all	20	48	30	30	Lateral aggregation	no exits	1	0
M1-3	CNV time-space	all	20	48	15	15	Lateral aggregation	no exits	1	0
M1-4	CNV time-space	*ongeval + gestrande*	20	48	15	15	Lateral aggregation	no exits	1	0
M1-5	CNV time-space	ongeval	20	48	15	15	Lateral aggregation	no exits	1	0
M1-6	CNV time-space	all	20	24	120	120	Lateral aggregation	no exits	1	0
M1-7	CNV time-space	all	20	24	30	30	Lateral aggregation	no exits	1	0
M2-1	CNV time-space	*ongeval + gestrande*	20	30	30	30	Full resolution	exits	1	10
M2-2	CNV time-space	*ongeval + gestrande*	20	30	30	30	Full resolution	exits	1	10
M2-3	CNV time-space	*ongeval + gestrande*	20	30	60	60	Full resolution	exits	1	10
M2-4	CNV time	*ongeval + gestrande*	20	30	60	60	Full resolution	exits	1	10
M2-5	CNV time	*ongeval + gestrande*	20	30	30	30	Lateral aggregation	exits	1	0
M2-6	CNV time	*ongeval + gestrande*	20	30	30	30	Lateral aggregation	exits	1	10
M2-7	CNV time	*ongeval + gestrande*	20	30	30	30	Full resolution	exits	1	0
M2-8	CNV time	*ongeval*	20	30	30	30	Lateral aggregation	exits	1	10
M3-1	FC	*ongeval*	20	30	30	30	Lateral aggregation	no exits	1	0
M3-2	FC	*ongeval*	20	30	30	30	Lateral aggregation	no exits	1	0
M3-3	FC	*ongeval*	20	30	30	30	Lateral aggregation	exits	1	10

Interpretation of results

For each test area, the NN is trained on data for 20 months and on a basis of a set of incidents. For *ongevallen* the number of incidents is in the range of a few hundred (depending on the length of the road section selected for the tests). About 25% are set apart for testing: this means that we use a set of in the range of 200-300 incidents to train the network to detect traffic patterns associated to incidents.

Once trained the NN is applied to the Test set to detect if the NN is capable of distinguishing incident vs. no-incident situations. The validation is based on the comparison between the actual image label (incident or no incident) with the label predicted by the NN.

For each result we calculate:

- The probability for incident/no-incident as computed by the NN
- The probability of incident per hour of the day based only on past incident distribution in the same road section in the same period of 20 months
- In each case we use thresholds to calculate the total number of:
  - True positives: images are correctly predicted as “incident”
  - False positives: images are wrongly predicted as “incident”
  - True negatives: images are correctly predicted as “no incident”
  - False negatives: images are wrongly predicted as “no incident”
- This lead to the ROC curve that measures the uplift of the NN compared to only statistical data

Result M3-2

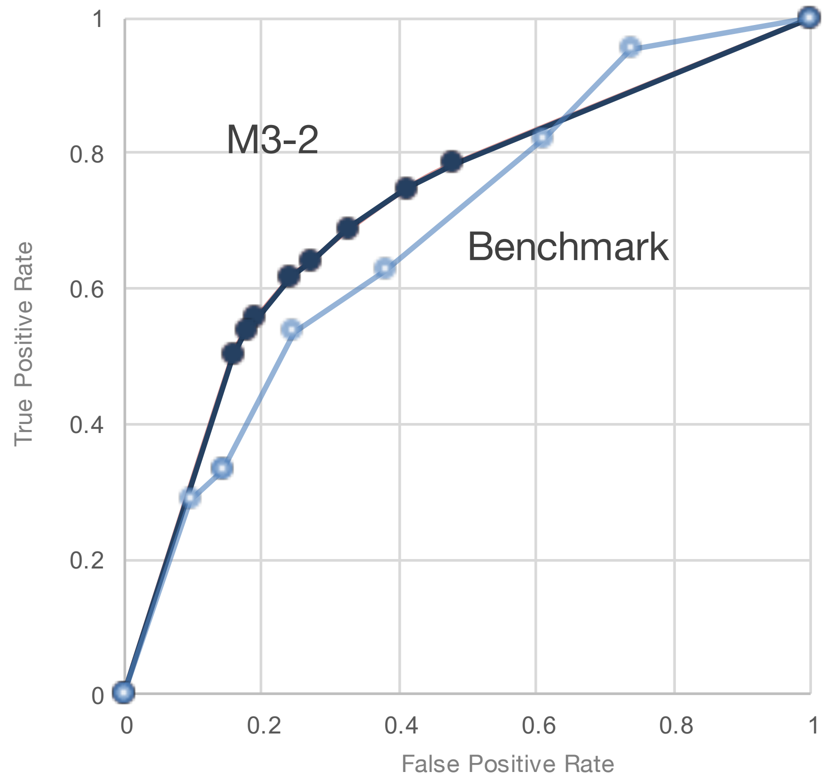
Incident set A12 Re, September-December 2014.
Total number of images considered: 81,180
Images tagged as “incident”: 1,630

Confusion matrix M3-2 with threshold 0.3

	T1	T0
Pred1	966	20833
Pred 0	601	64780

Confusion matrix benchmark

	T1	T0
Pred1	983	32667
Pred 0	584	52946



Notes:

- The confusion matrix for the benchmark is presented for a threshold that gives ± the same true positives compared to M3-2.
- M3-2 produces ± the same true positives and false positives, with about 1/3 less false negatives and about 20% more true negatives
- Note that this applies to traffic images (one image per minute in best cases) and not on events, which usually span multiple images sequentially.

Result M3-3

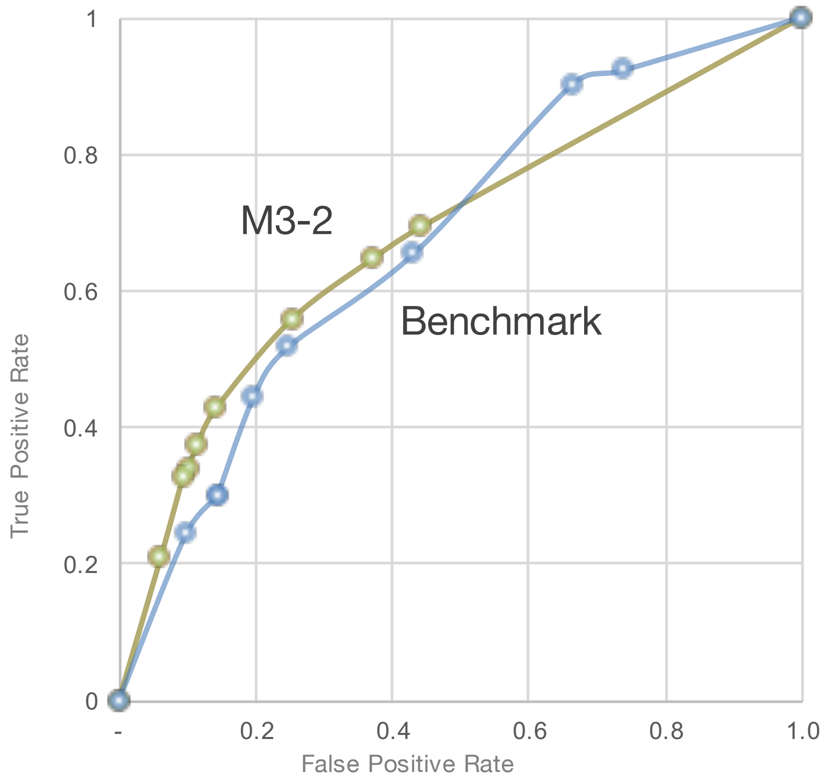
Incident set A12 Re, September-December 2014.
Total number of images considered: 81,180
Images tagged as “incident”: 1,630

Confusion matrix M3-3 with threshold 0.1

	T1	T0
Pred1	754	13744
Pred 0	876	71806

Confusion matrix benchmark

	T1	T0
Pred1	723	16888
Pred 0	907	68662



Notes:
- The confusion matrix for the benchmark is presented for a threshold that gives ± the same true positives compared to M3-3.
- M3-3 produces ± the same true positives and false positives, with about 20% less false negatives and more true negatives
- Note that this applies to traffic images (one image per minute in best cases) and not on events, which usually span multiple images sequentially.

Result M1-5

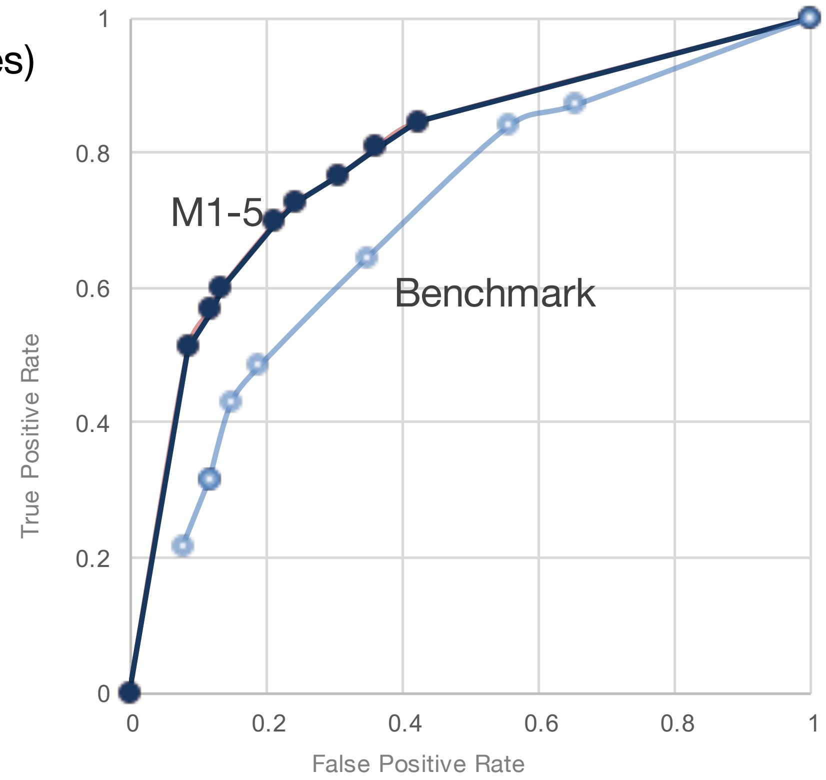
Incident set A12 Re, September-December 2014.
Down sampled set (67% no-incident – 33% incident images)
Total: 2600 images

Confusion matrix M1-5 with threshold 0.5

	T1	T0
Pred1	553	268
Pred 0	313	1466

Confusion matrix benchmark

	T1	T0
Pred1	529	545
Pred 0	339	1189



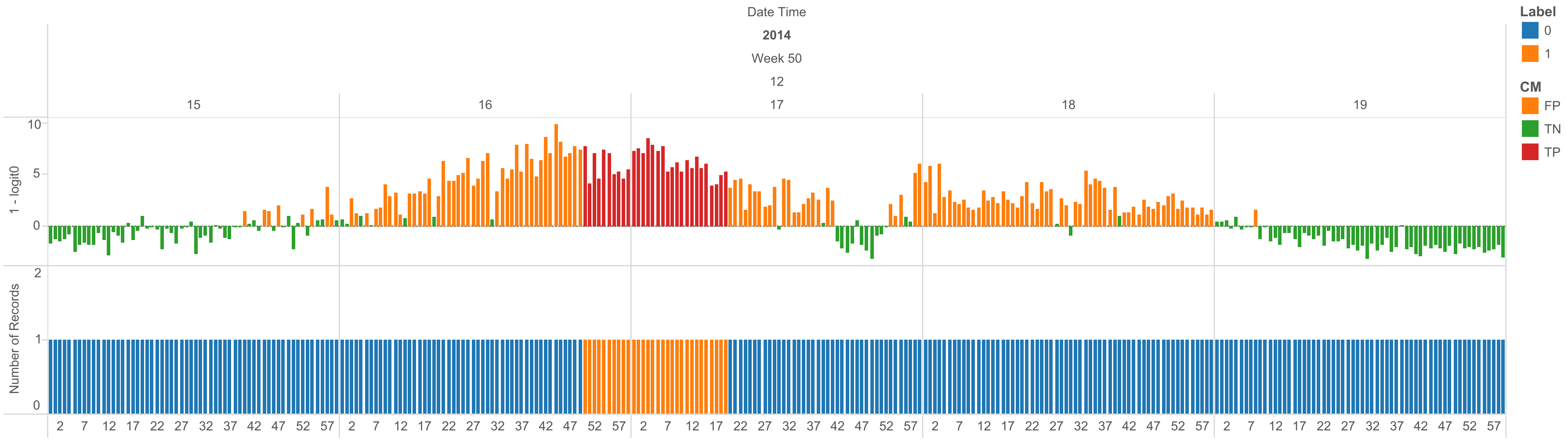
Notes:

- The confusion matrix for the benchmark is based on a threshold that gives the ± same true positives compared to M1-5.
- M1-5 produces the same true positives with about ½ false positives, ± same false negatives and significantly more true negatives.
- Note that this applies to traffic images (one image per minute in best cases) and not on events, which usually span multiple images sequentially.

Details of incident identification: example 1

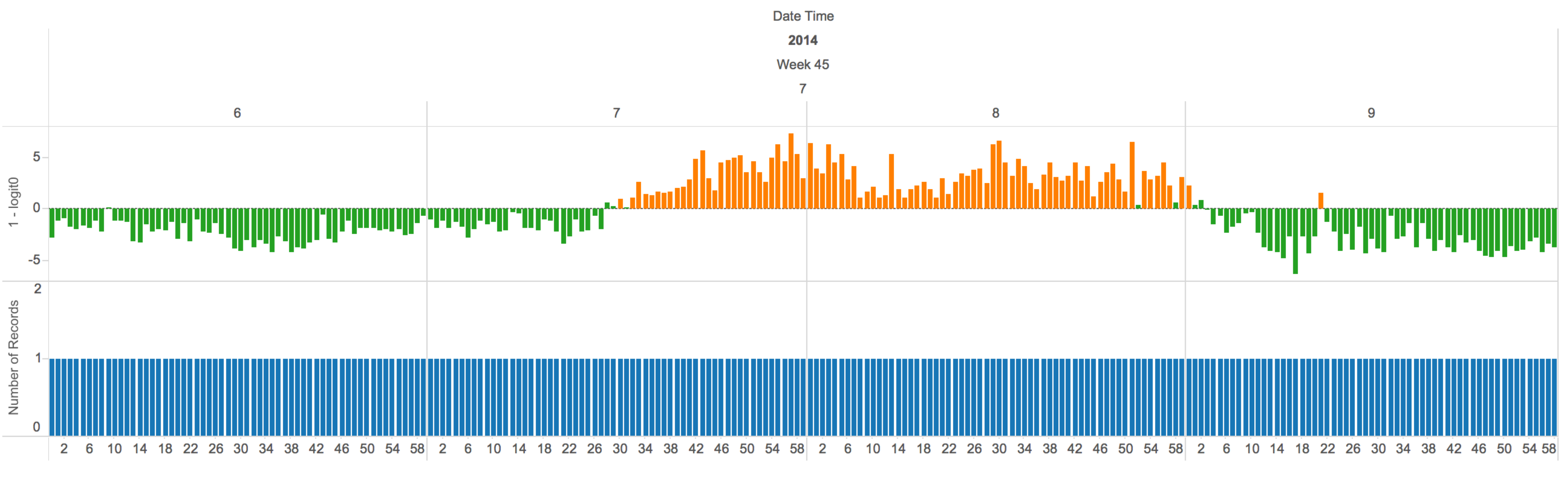
The example plots the incident probability (more precisely, the 1- logit_0 function, top diagram) and compares with the event timeline (bottom diagram). The event diagram shows in orange the first 30 minutes after an incident was recorder (based on the VIAS recorded start time of the incident).

The NN indicates a higher and sustained probability of incident in the 30 minutes before the incident.



Details of incident identification: example 2

The example plots the incident probability (more precisely, the 1- logit_0 function, top diagram) and compares with the event timeline (bottom diagram). In this case, the traffic patterns were compatible with higher incident probability before and around 7am and around 8am, but there was no incident.



4

DISCUSSION AND CONCLUSIONS

Summary

- The project tested multiple types of Deep Learning Neural Networks applied to the prediction of incidents on the Dutch highways.
- With the term "prediction" we mean predicting the status of a road section (S) in a time window (T) for a type of incident (I). With "status" we mean "Incident" or "no incident".
- Traffic data, measured by highway loops, has been used as the sole predictor of incidents. Patterns in traffic data during a time interval before the incident (H) are used as predictors.
- The project confirms that Deep Learning Neural Networks trained on past data are capable of increasing the predictability of incidents above pure historical statistics. This implies that the **NNs detect traffic patterns that precede incidents** and are – in part - able to single them out in space and time. This has been achieved by training the neural networks with **very small training sets**, of a few hundred incidents only!
- The models have been tested for multiple road sections and in particular for *ongevallen*.
- Considering that traffic is only one of the many determinants of incidents and that the data used has many limitations, the NNs have performed well.
- It is realistic to expect that better traffic input data combined with additional sources of data and further tuning of algorithms will lead to significant increases in incident predictability, to the point where it can be used as a source of insights for traffic management in near real-time.

Main conclusions

The NNs can uplift prediction accuracy on a training set of ± 300 incidents over 2 years. The size of the training set and data quality appear as the main limit to the ability of training for shorter sections of the highway.

We have evidence that useful results can be obtained for a wide range of time windows in the future (from 15-30 minutes to 2 hours). This indicates that the methods can be developed for a wide range of operational situations.

The models have been trained on a 2-year dataset. The 2 years choice is dictated by data availability but it seems reasonable as traffic patterns appear to change in the 2 year period.

Data incompleteness and/or inconsistencies are a major determinant of the overall results. Without improving data completeness/consistency it is difficult to achieve the potential uplift of the NN models.

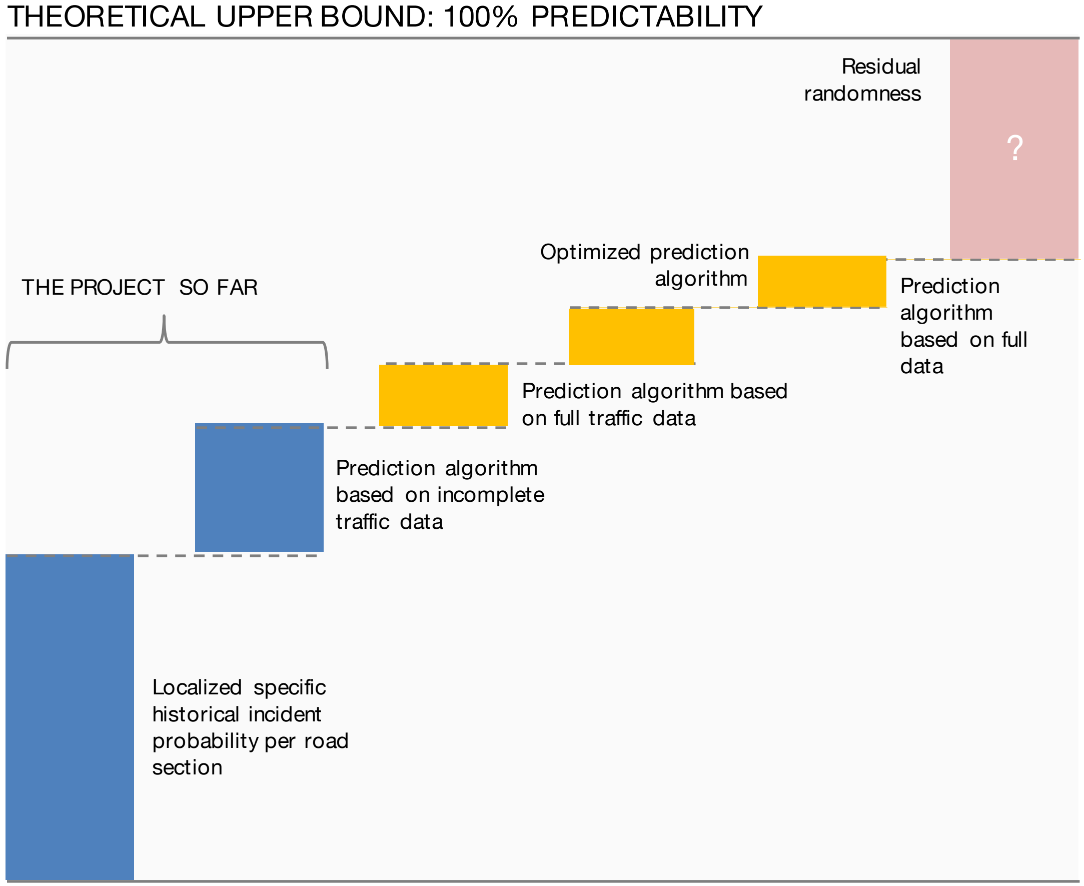
We tested about 50 models, broadly clustered into five classes of parameters and using two types of NNs: Fully Connected and Convolutional. Both types perform well. However, the general assessment is that FC can be applied across all situations, while CNV could be superior within specific local settings.

Main observations and learnings

The following are the main observations and learnings from the project (see details on the next pages):

- Uplift progression. NN learns patterns on a small number of incidents: this points towards a natural uplift progression to get to reliable operational predictions.
- Data quality and completeness. This is the strongest barrier to improving prediction capabilities.
- The prediction canvas is wide. NNs have proven able to cover a large span of the prediction canvas.
- Scaling the models may require a mix of strategies.

Uplift progression



The key premise of the project is to validate the ability of machine learning to produce superior insights compared to statistics alone. This has been proven.

While 100% predictability is unrealistic, three questions lead the evolution of these methods:

- Which are the measures that improve the quality of predictions?
- Which “prediction” and which “quality” provide most value to Incident Management ?
- What is a technically attainable goal and which is a stretch goal ? (for instance: 50% incident prediction with 25% false negatives.)

Data quality and completeness

There are several issues with the data that impact the ability of training the NN, but two in particular are crucial:

- The traffic data is incomplete at every single location and point in time. Since the model is based on pattern recognition that scans sequential images, sequential completeness is also an issue. All in all, the project managed to use ± 50% of the available data by accepting between 80% and 90% of data completeness, depending on the tests.
- The incident data includes crucial information (e.g. start time of the incident) to correlate traffic and incidents. In several cases there are legitimate doubts about the accuracy of incident time registration.

The implications are multiple:

- A much smaller number of viable data sets to train the networks, which impacts especially short sections and short time frames
- A high degree of noise in the data, which confuses the network and degrades predictions
- A significant learning curve to detect signal patterns in the data
- An overall lower ability to predict

Addressing data quality and completeness

The data completeness issue was known to previous projects of Incident prediction. However, in this project we have explored the issue in great details. There are few viable options to address the issue:

- **Data imputation/interpolation**: a process of imputing the missing data from the existing data so that the dataset increases its completeness to the level where NN can effectively learn patterns. This has been tested and for small gaps it is a viable solution, but for larger gaps it is too coarse and does not produce benefits.
- **Image interpolation and processing**. Since the traffic (and other data) is structured into images, tools from remote sensing can be used to smooth out images and fill gaps through spatial interpolation.
- **Data remodeling.** Introduction of an abstract model, where data for each road section is derived from multiple sources (loops, but also other such as floating car data) and processed in a way that changes to the road usage (e.g. lane closure) or data feed (e.g. loop removal) does not imping the NN structure, which remains the same.
- **Data prediction**. In the same fashion used to predict incidents, it is possible to train an algorithm to learn from patterns of traffic and fill gaps based on the pre-post gap data available and impute patterns in between that are more appropriate than simple interpolation alone.

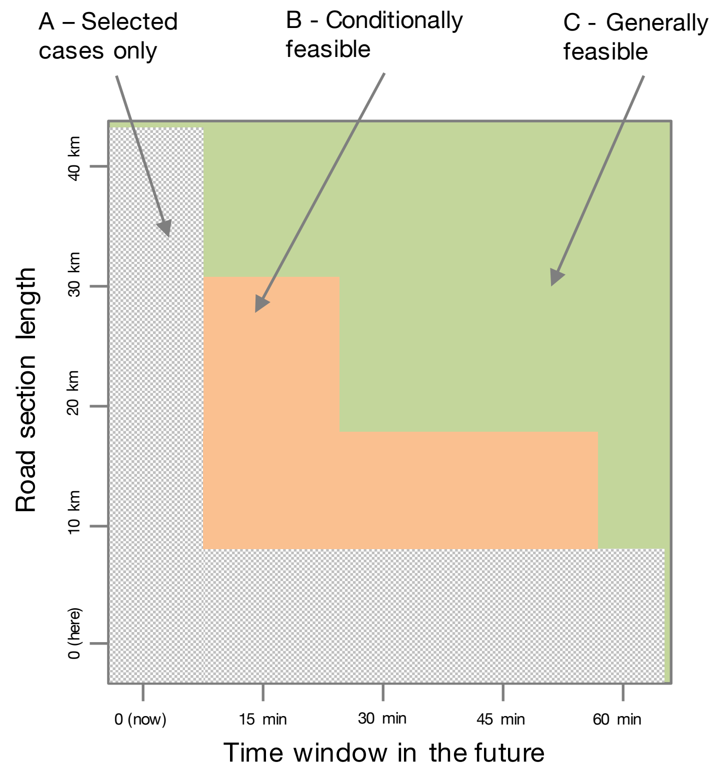
Prediction canvas

The prediction canvass describes combinations of road length and time window in the future for predictions. As a broad indication, the canvass illustrates a few regions with the following characteristics:

A: This area has narrow time and space constraints, such as ”Predict incident in a section of 2km in the next 5 minutes”. Feasibility requires localized models and sufficient training data. This is realistic only in selected places, but not at scale.

B: This area has either a time or a space criticality, which can be addressed through longer training histories. Alternatively, this could be a candidate for transfer learning, which would make training transferrable across the network (thus increasing the options for training).

C: This area is where the current models have been tested to work, and for which there is sufficient insight to confirm that the models can be improved with additional sources and longer histories to obtain sufficiently accurate predictions.



* The thresholds in time and space in the canvass are for indication and should be interpreted as fuzzy margins.

Scaling the model

So far we have tested the models on sections of A12 and A10/A9/A4. The tests included main carriages with and without exits. The prediction models have been adapted to each section considered to use the input available for each section (# of loops, configuration etc.) but also to accommodate to diversity in the data available.

This approach is however not scalable. Scaling the model requires two additional steps:

- Data abstraction. By separating the prediction model input from the data input through an intermediate data abstraction model we can accommodate input variations (on a single section of the road but also across multiple sections) as well as input differences.
- Transfer learning. An assumption of traditional machine learning is the training data and testing data are taken from the same domain (we train the NN on A12 and test on A12). Transfer learning refers to high-performance learners trained with a certain dataset in a domain (e.g. A12) and then applied to a different one (e.g. A4). This can be used to substitute a training set or to complement a training set.
- We have tested transfer learning on the existing models and it performs modestly. The reason is most likely the strong relationship between model and road section. With data abstraction, this dependence should weaken significantly and it is likely that transfer learning will to some degree work. The major benefit would be the increase of training data and thus the direct impact on quality of predictions.

Next steps

There are many options to develop these methods further with the goal of assisting traffic managers dealing with incidents. This is a high-impact exercise with many consequences, and it requires multiple iterations to bring the full results to fruition. From the current stage, the following are the most immediate opportunities for improving results:

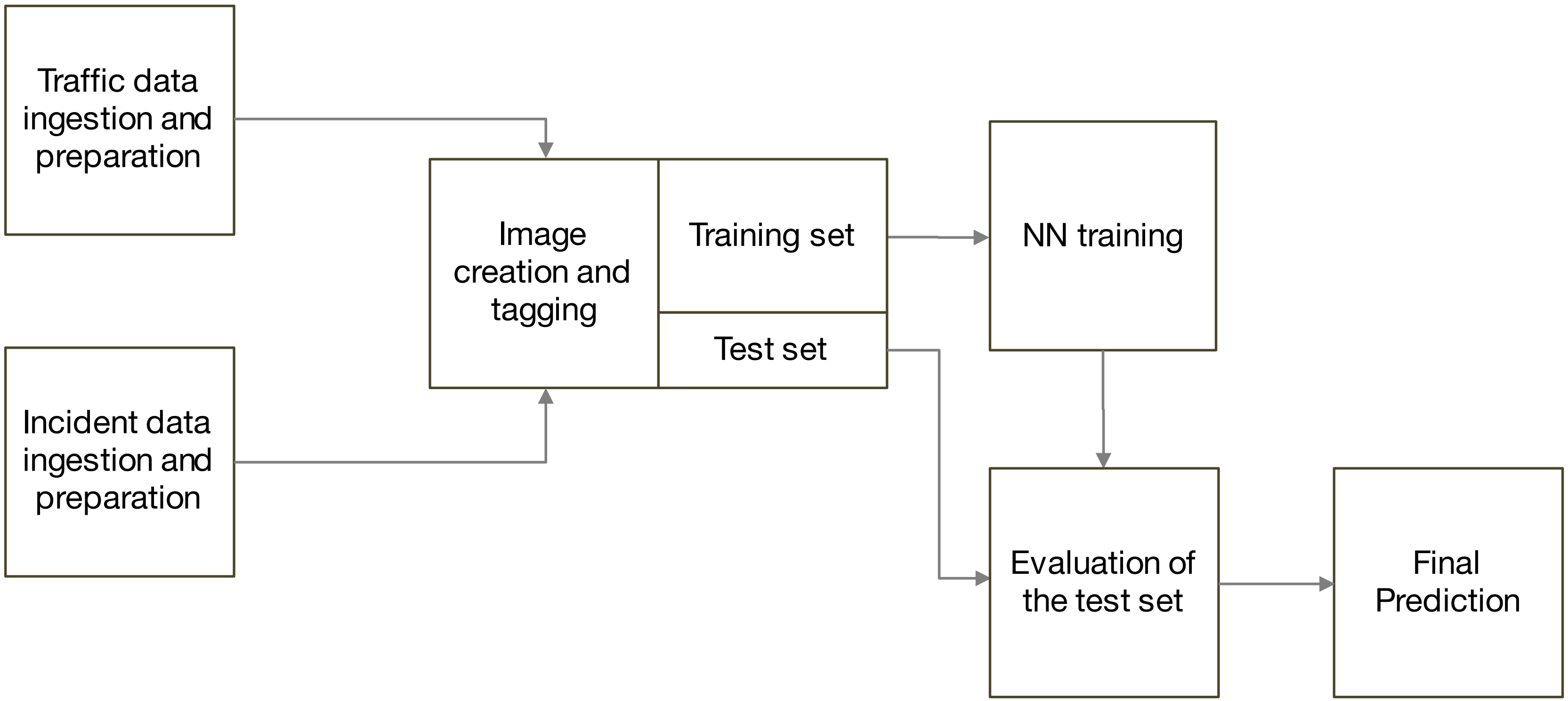
- Data related actions
  - Improving data quality through interpolation/imputation and reducing ambiguity/sensitivity on incident classification and time registrations, possibly through LSTM algorithms.
  - Data generalization and abstraction, to separate prediction model from data input, possibly introducing floating car data as input.
  - Improve and validating the quality of incident data through direct review and blending of multiple datasets and machine interpretation of the data.
- Data set additions
  - Adding highly localized weather and local environmental data.
- Interpretation
  - Iprove the result interpretation and validly by visualizing the features extracted by the hidden layer of the NN to aid interpretation of the rationale of predictions
- Scalability
  - Introduce transfer learning (in combination with data generalization and abstraction) and test scaling through transfer learning.

A

APPENDIX: IMPLEMENTATION NOTES

Workflow

A high-level description of the workflow used in the project.



Traffic data set ingestion and preparation

The loop datasets, provided by NDW, contains three types of data relevant to the study:

- Loop measurements per 1 minute. We use complete records for the years 2013-2014.
  - Loop id: the unique identifier of the loop in the road network
  - Index of lane number: the lane numbering, starting with # for the left-most lane
  - Vehicle speed (km/h) -1 minute, Car flow (# of vehicles in the last min) - 1 minute
- Lane index lookup table
  - indicates the distinction between lanes for loops
- Loop metadata
  - loop locations plus other loop metadata
- Data selection traffic
  - From CSV we keep flow and speed for category “any vehicle”. Depending on the model used we select data from the main carriage way including or excluding the exits
- Data pre-processing traffic
  - If two loops have different ID but same location, they are merged: result in loop Gates
  - Data is split per direction of travel and selected for the test road section

Incident data set ingestion and preparation

Incident data is provided by RWS. The VIAS data set is used in the project:

- Data is available for 2010-2016 (registration methods have changed in the period)
- Incident data selection
  - Only incidents with a valid location, valid start time and valid classification are retained
  - Incidents are then selected for test road section
  - For incident time we use the recorded start time

Image creation and tagging

- The data are processed with python scripts and turned into the form of “images”. The image structure is as follows
  - Columns = location
    - If the data is aggregated across lanes, one location per loop-lane set
    - If the data is non aggregated then one value per loop-lane
  - Row = traffic values per pixel (encoded as pixel color)
- The script does several processing steps in sequence:
  - Data selection (selection of loops of the road segment in question for the time period of the training)
  - Lateral aggregation (averaging over lanes, if applicable)
  - Image creation with split between TRAIN ± 80 % of images) and TEST (rest), TRAIN and TEST are sequential slots of images.
  - Image filtering depending on “completeness threshold” (usually 90%)
  - Image tagging buffer creation (10 minutes): an image is tagged as “incident” or no ”incident” if in the following time window (T) there is (there isn’t) an incident in segment S. A buffer of 10 minutes is added to compensate for possible incident time registration errors.
  - Uploading the both datasets to S3 bucket

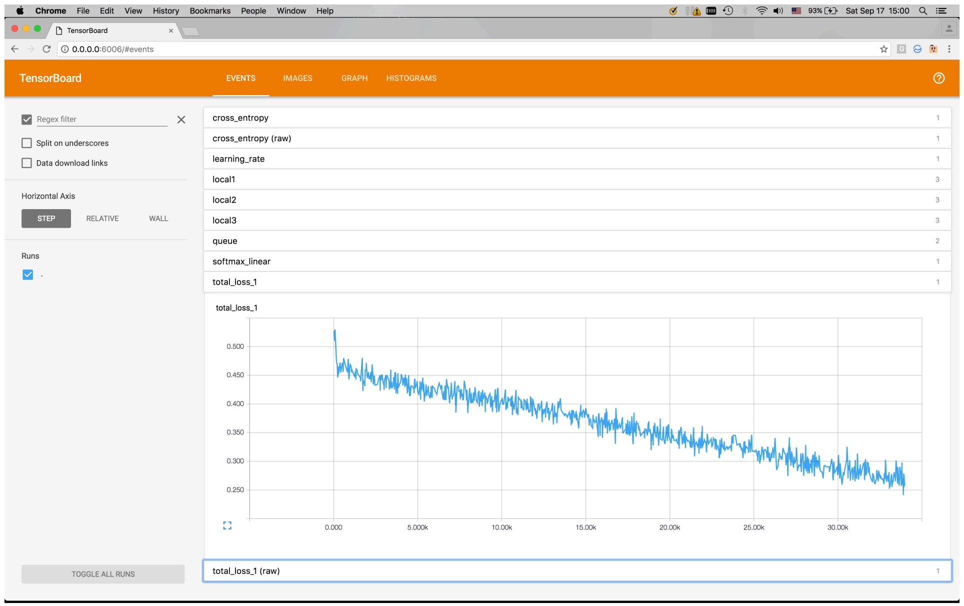
	L1	L2	L3			
	Speed Flow					
	…	…	…	…	…	…
		Speed Flow	Speed Flow	Speed Flow	Speed Flow	
T0 +2D						
T0 +D						
T0						Speed Flow

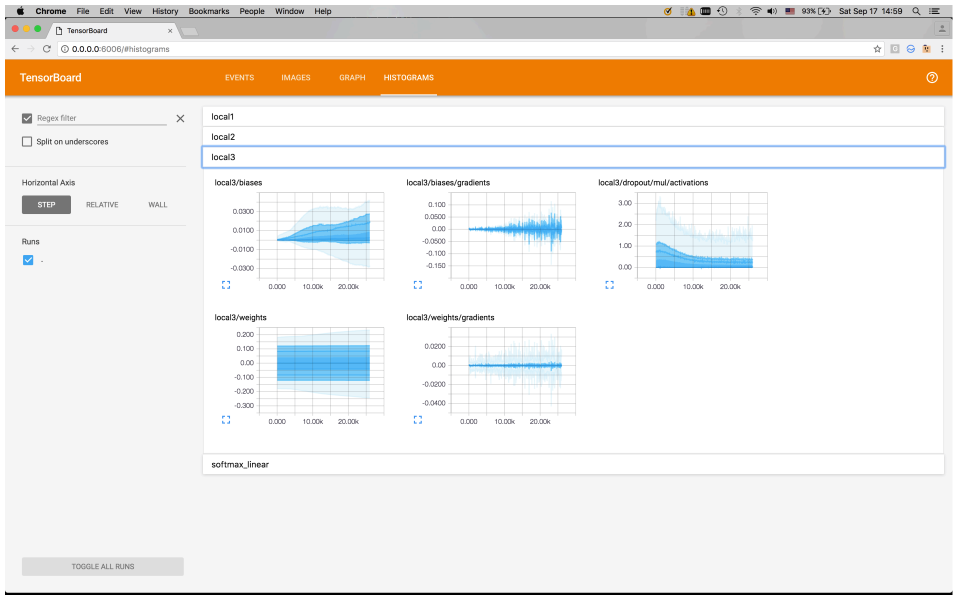
Example of image (D = time delta, e.g. 1m, 10m)

Note: We used a native Tensor Flow file format for storing the images – TFRecords

NN training

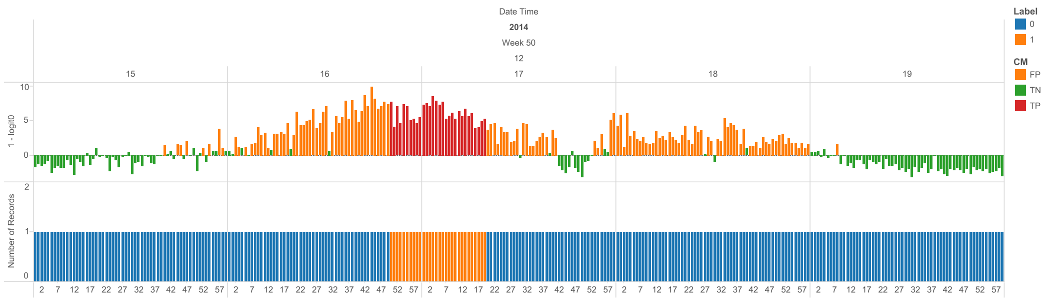
- Neural network training tales place either on the AWS or Rescale platforms (see Tools and infrastructure slide).
- Neural Network training can take a significant amount of time (can range between 2 and 24 hours in most practical case).
- During the training the processed is monitored through specific tools (e.g. Tensor Board, see below).



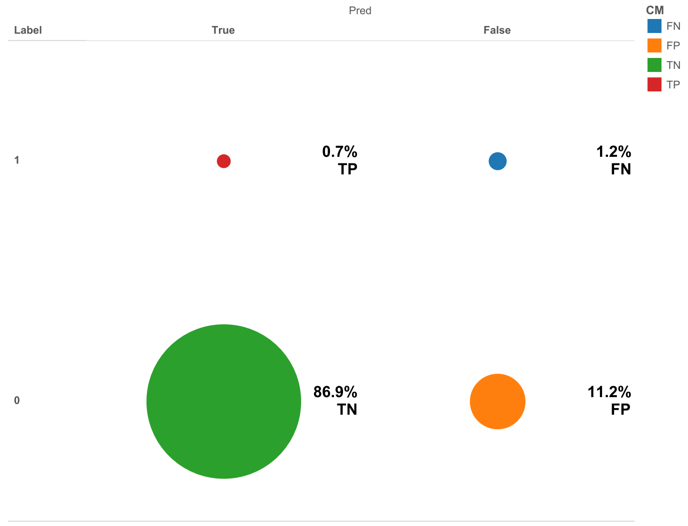


Evaluation of the test set

- When applied to the test set, the network outputs the predictions in binary format and Logit probabilities.
- The final predictions are presented in the form of the Confusion Matrix where we compare the image tags and the neural network predictions
- One image may correspond to several incidents, and an incident labels many images.



Example of probability outputs (logit) for the NN. The orange block indicates that the NN favours "incident" prediction. Red block in the function means a True Positive – correct precision od an incident



Example of confusion matrix.

Tools and infrastructure

rescale

amazon web services

AWS EC2 server + REDSHIFT DB + local machine:

GIS pre-processing
Data Pre-possessing
Data evaluation
Plotting results

Data →
← Results

amazon web services

AWS S3 storage:

RAW data storage
Processed data storage
Results storage

Data →
← Results

NN training on GPU arrays

Tensor flow on Rescale platform

Data →
← Results

NN training on CPU arrays

AMAZON EC2 platform:
High performance GPU computing

amazon web services

Storage:
- AWS S3 (storage): 1TB
- AWS Disk for data processing in EC2: 8TB temporary disk
- AWS Redshift (data warehouse): 1TB

Servers
- General purpose server for processing: AWS EC2 server, Linux, 30 GB RAM, 8 CPU cores
- Server for training Neural Networks: AWS EC2 server, Linux, 32 cores, 244 GB RAM, 1TB Disk; Software installed: Python, Theano, Tensorflow
- Rescale platform: 4 GPU arrays each 4GB memory – NVIDIA GK104, Software installed: Tensor Flow, Storage: up to 100GB

TensorFlow

Open source software library for machine learning in various kinds of perceptual and language understanding task.

theano

A Python framework for fast computation of mathematical expressions with multidimensional arrays.