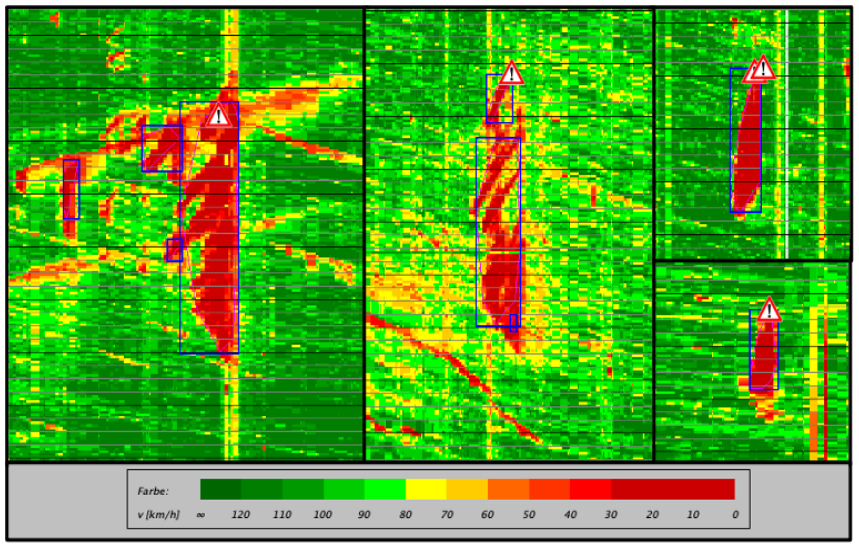
The search for predictability and correlations between traffic-data based congestion- and incident characteristics – An exploratory data analysis

Master‘s Thesis by Jakob Erpf

Mentors:
M. Eng. Barbara Karl
Dr. - Ing. Matthias Spangler

External Mentors:
Dipl. - Ing. Stefan Gürtler (S&W)
Dipl. - Ing. Johannes Grötsch (LBD)



As the title describes, the thesis’s goal was to identify correlated features of jams and incidents. This was approached by an exploratory data analysis of three real world datasets containing traffic movement, accident and roadwork data.

The first step of the methodology is the jam detection based on the traffic movement data (floating-car-data). This was done by implementing a performance tuned version of the density-based cluster algorithm DB-SCAN. After the detection of jams, the incidents in form of accidents and roadworks are matched to temporal and spatial adjacent jams.

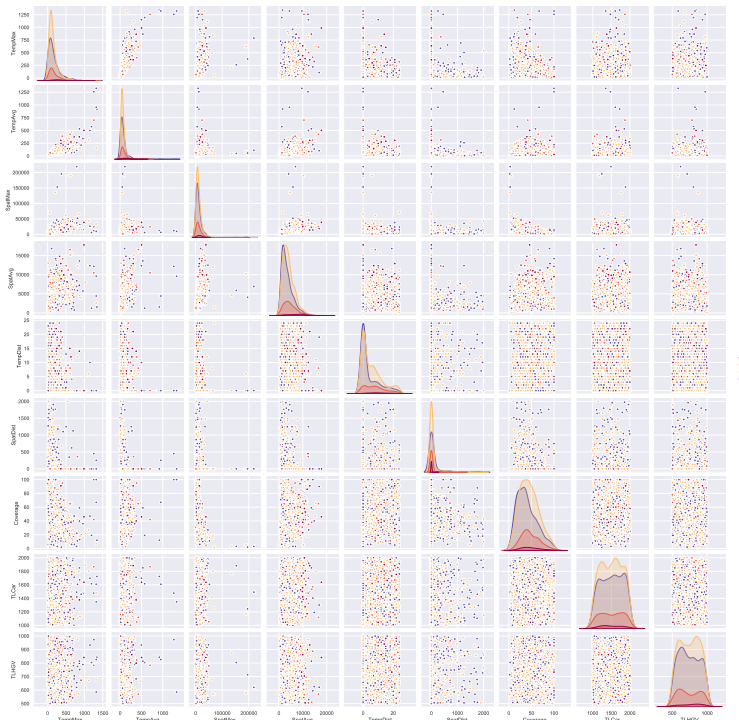
This algorithmic processing of the dataset results in congestion – accident/roadwork matched which are exemplary show in the figure to the left.

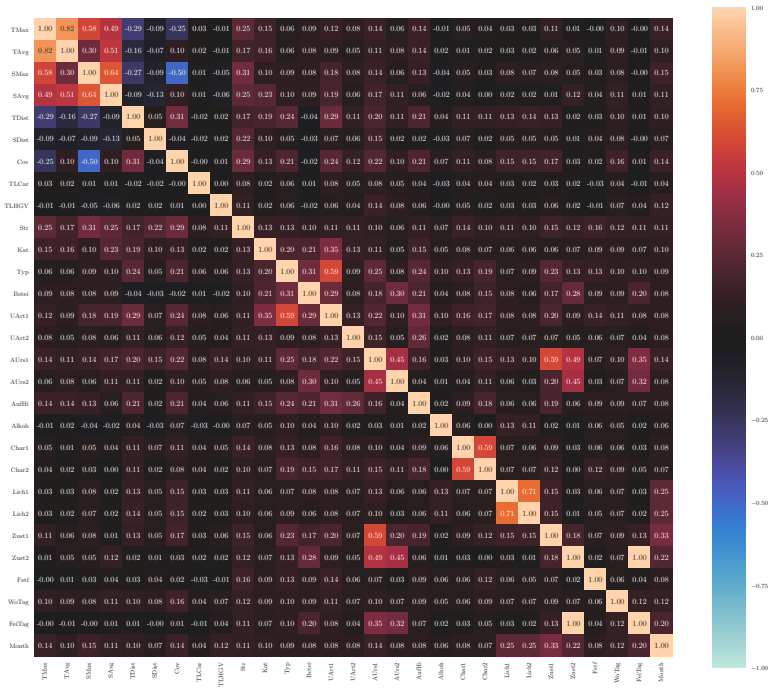
This jam detection and incident matching results in two datasets which contain a list of congestion – accident and congestion - roadwork matches which are the analyzes for relations.

By using common data analytic framework like Pandas and NumPy the datasets where initially reviewed visually to identify possible relations and provide an overview of the datasets. After defining the attributes to be research from visual representations like in the figure on the right, a deep dive into statistical mathematics it taken, to identify and define the statistical methods to be used for proving association in mixed datasets. This results in five correlation coefficients which are suitable for the analysis of each attribute relation found in the datasets and other statistical methods for proving significance of the relations.

These coefficients and methods are implemented in a combined Python and R processing tool which computes the correlation and significance of all relations with the suitable statistical method. This processing results in a correlation matrix (exemplary shown in the figure below) and significance matrix which can be used to visually identify significant relations.

These statistically significant relations are then further analyzed with a pairwise Kruskal-Wallis test to find interpretable differences in the variables.





The result of the analysis showed that there are characteristics of jams and incidents which are significantly related. Among other effects this means that the accident category, type, as well as the month and the month and the location (road) statistically indicates the length and duration of a jam. The coverage, describing the density of a jam is also correlated with many accident attributes.

Through the analysis of the different categories, referenced by the attributes in the datasets, it also became clear that many categories have an insufficient sample size for significant results which presents the need for further research with larger datasets.

The separate analysis for predictability showed that although there are significant correlations, these relations do not have any predictability associated with them. As a result in can be stated that congestion and incident characteristics significantly correlate with each other but do not provide predictability, based on the provided datasets and applied methods.

The complete thesis with scripts, plots and results is available in the GitHub repository reachable under: https://github.com/jakoberpf/master-thesis