Accident Analysis in Zurich

Research Proposal Final Project
ESC403 Introduction to Data Science

Marcelo Looser, Dave Linder, Yves Bicker

May 6, 2021

We want to analyze what circumstances lead to accidents in Zurich and if we can project a heat map displaying the probability of high risk areas by means of finding what features are the most important to predict accidents.

What questions will you try to answer with the project?

- Correlation, causation?

  Performing statistical inference on the given data set will be used to get a grasp why certain accidents occurred (the causes might be very trivial). Once these are established, preventative measures will be proposed if there are any.

- Prediction, prevention?

  After establishing a sound basis why accidents occur, day time, weather and other features will be used to predict high risk areas for accidents in the future (this might also be trivial, high rate of accidents $\Rightarrow$ high risk area).

- Just for fun.

  If by chance certain correlations arise that clearly do not imply causation, it will surely be exploited for comedic relief.

What data will you be using (and what is the source of that data)?

We will use the road accident data of Zurich, the corresponding weather data and the Geo data catalogue, and road condition data if there is any available.

Links

Here are links to our data sources:

- Accidents - Data set
- Geodata catalog
- Meteodata

How will data be processed?

Once the data has been cleaned and checked for any anomalies, exploratory data analysis will be used to search for trends and features with the most influence. Afterwards, as mentioned above, we will use inferential as well as predictive analysis.

What analysis techniques or algorithms are planned to be used?

We want to use a clustering algorithm like PCA, or if this isn't fruitful sensitivity analysis, to determine the significance of the features used. Once the most significant features are determined, linear regression or some other interpolation techniques like spline interpolation or radial kernel interpolation or possibly nearest neighbour interpolation will be used to see the relationships between the features and the data points, either in 2 dimensions or higher depending on the data. Afterwards we will try to estimate a probability density function in dependence of the most influential features to get a heat map of the high risk areas where accident occurrences is most probable. This might be done through a small neural network or with more basic tools.

Snippets of the features in the corresponding data sets:

- Accidents - Data set
  - Accident type
  - Accident severity category
  - Vehicles and pedestrians involved
  - Road type
  - etc.
- Geodata catalog
  - Street noise level
  - Heat load on streets and vicinity
  - Urbanisation
  - etc.
- Meteodata
  - Temperature
  - Air velocity and orientation
  - Rain duration
  - CO concentration
  - etc.

  This features may or may not be used.