Exploratory analysis

Before the technical tariff structure can be constructed, an exploratory analysis of the given data has to be done. The exploratory analysis will be split up in two parts. The first part will explore the data for the frequency information and the second part will explore the data for the severity information.

Frequency

To get a feel for the data some basic R functions (str, names, head, summary) are used on the dataset “Data” with all the data and the dataset “data\_train” with just the training data. There are noteworthy findings after running these functions. Only some variables where you may expect type long are of the type character, like long, lat. This should be kept in mind when using these variables. After inspecting the data, the variables lnexpo and freq\_ann are removed from the dataset. They were considered irrelevant. ( extra uitleg)

To obtain a further understanding of the dataset a grid of relative frequency plots is made using the training data. A lot of useful findings can be extracted from this grid. Most policies In the dataset have 0 claims and few have 3 or more claims. There aren’t many cars that are that are younger than 1 year. The age of the policy holders is fairly even distributed. Most policies are held by males and have the standard MTPL coverage. Last there aren’t many cars with more than 110 horsepower or are a sportscar, that are part of a fleet or have a professional use.

(pic: exploratory relative frequency)

To make sure the training dataset “data\_train” is a good sample for the full dataset “Data”. The empirical frequency of both datasets is compared. The empirical mean and variance for the whole datasets are almost the same. The empirical variance of the individual variables is also compared. From comparing these empirical variances there can be concluded, that the training dataset is a good sample for the actual full dataset. (make table with numbers maybe?)

Severity

A good feel for the data has already been developed during the frequency analysis. So there can immediately be looked at the density of the claim amounts. From the graph can be conclude that most claim amounts are grouped together. There are two peaks in the claim amounts as can be seen in the graph.

(graph density claimAmt)

A map of Belgium is made which displays the relative claim amount per region. This map gives a good overview in which regions the highest claim amounts occur. The most notable region with high claim amounts is clearly Brussels and the surrounding regions.

(MTPL claimamount map Beglium)