## Binning of spatial data

Spatial data was represented as a pair of coordinates in the dataset. These corespond to the center of each of the Belgian cities. In this way they could be linked to the postal codes, which is a known variable for each policy holder in the dataset. This allows us to make predictions of the claim frequency and claim severity based on the city that a policy holder lives in. Then based on these predictions the postal codes can be binned in an optimal number of factors.

To start, a model to estimate the predicted frequency and severity needs to be constructed. A Generalized Additive Model will be used for this purpose. A basemodel to model spatial data is y ~ s(long, lat, bs=”tp”). From here on, there will be differences between frequency and severity because they will each have their own optimal GAM-model, based on different dataset. Frequency used all observations, while the severity dataset makes abstraction of policy holders that did not file a claim in the observation period. Note that a smoother was used for longitude and latitude coordinates, which makes us able to see regions with higher expected claim frequency or expected claim amount. If only postal codes were used, these regional effect would not be visible and all cities would be seen as independent of their location with respect to each other.

For frequency, we use a Poisson-family with a log-link function which is a logical choice when modeling claim frequency in an insurance context. Also, the exposure needs to be added as an offset due to a not all policy holders being covered for a full year and due to modifications (eg. people that move change their postal code, people that buy a more powerfull car). The spatial model is then extended with different other variates. By comparing the log-likelihood and the Akaiki Information Criterion (AIC) of different GAM-models, the most optimal GAM for frequency data was found: freq ~ s(ageph) + s(long,lat) + power + cover + fleet + split + fuel + sexph + agecar. Also, this model was constructed with a Restricted Maximum Likelihood (REML), while still applying the offset and the same family and linkfunction.

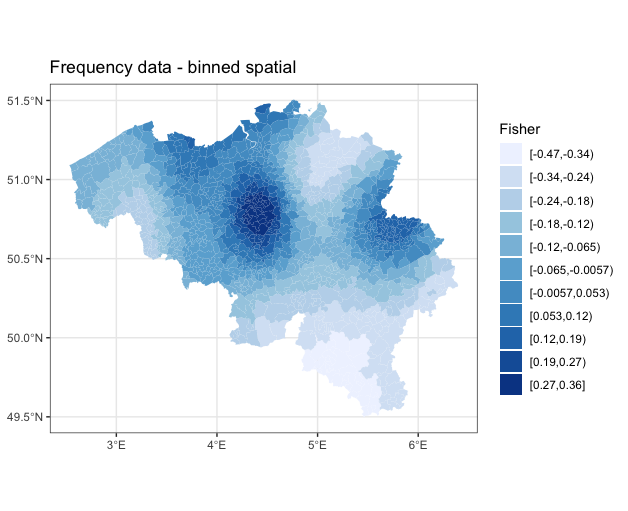
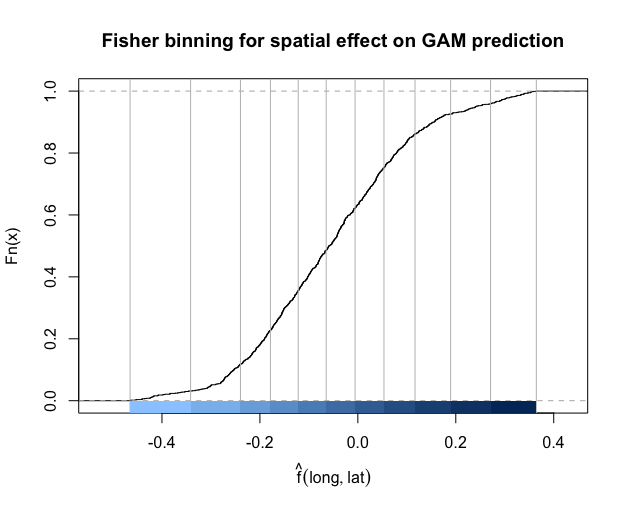
For severity …

These models were then used to predict the expected claim frequency and amount per city. On a continuous scale, this would yield results presented in the following map of Belgium:Afbeelding met kaart

Automatisch gegenereerde beschrijving

Linking each policy holder’s postal code to its predicted frequency and amount, makes us able to bin the continuous variable using Fisher’s method. This method bins based on the steepness of the cumulative distribution function of the predicted values.

A follow-up problem of binning is choosing the optimal number of bins over which the observations need to be distributed. By comparing the BIC’s, for frequency modelling 11 bins came out to be optimal. This yields a factor variable that approximates the continuous spatial variables. The factor variable is added to the dataset to be used in further modelling.



For severity … bins were found to be optimal.

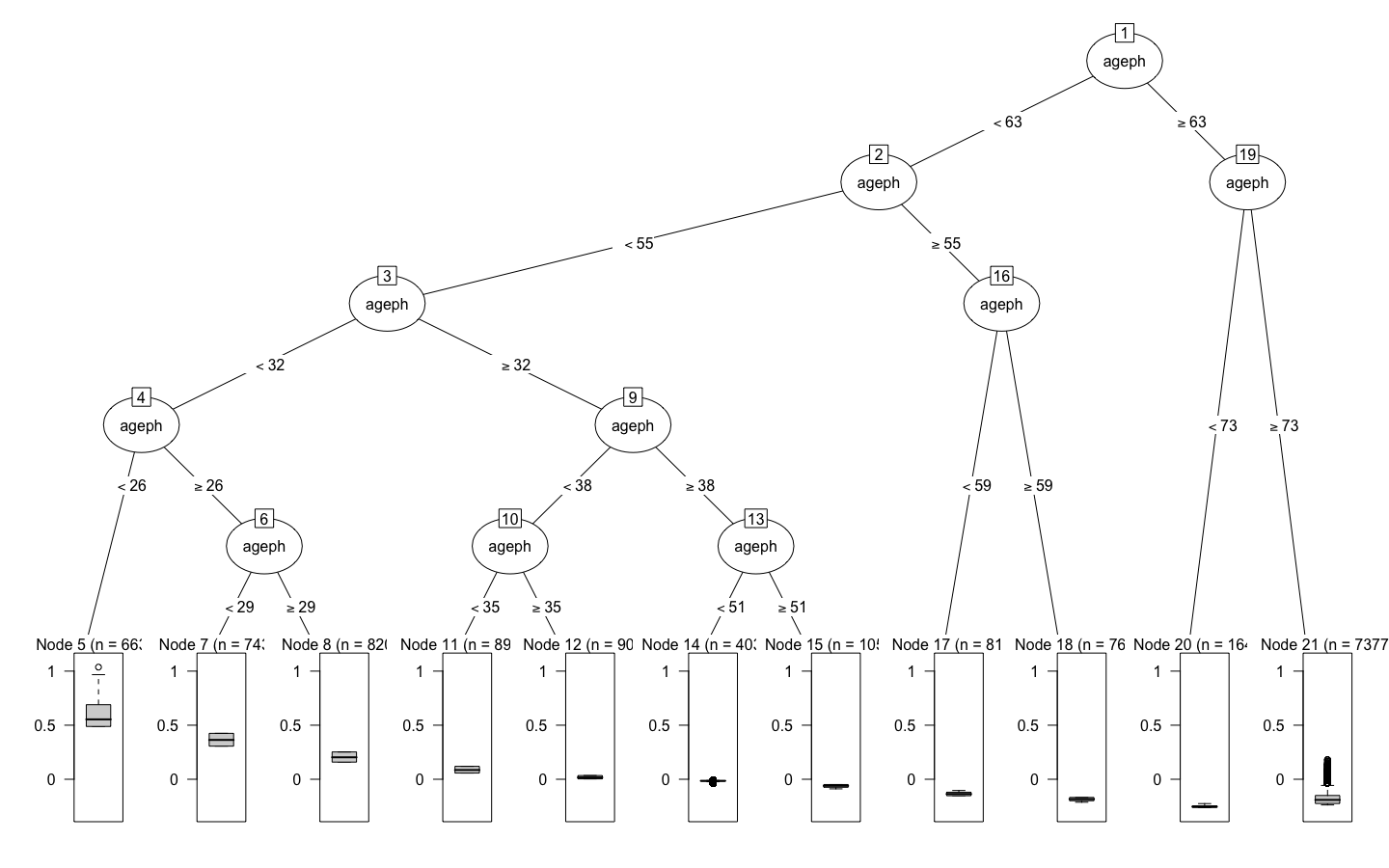
(afbeeldingen)

## Binning of age variable

In the dataset used to price our insurance products, age is recorded as a continuous variable with ages ranging from 17 to 95 years old. To optimally use a GLM, it is advised to use only categorical variables, and thus age is best converted to an ordered factor variable. Again, because there is a difference in the used dataset for frequency and severity modelling, the breakpoints may differ.

For claim frequency, the GAM-model used for the binning of spatial data was reused, but with the binned version of the spatial data as a replacement for its continuous counterpart. All other aspects identical. So the used formula is: freq ~ s(ageph) + geo + power + cover + fleet + split + fuel + sexph + agecar, with ‘geo’ being the binned spatial variable. By predicting the claim frequency and constructing a new dataset (GAM\_data) where the observations are counted per value of age. This is possible because ‘age’ is not really a continuous variable, it can only take positive integer values. In this dataset, the coefficient of the smoother of ‘age’ is also included. Based on GAM\_data, the evtree( )-function in R can be used to construct a regression tree based on an evolutionary algorithm and with the counts per age as weights. In this function, it is usefull to include an evtree.control( ), which controls the complexity of the constructed tree. 4 Parameters were used for this goal, with the first being the alpha, which was set to 100. Also, the maximum depth of the tree was set to 5 to control the size of the tree. The two last control parameters were set to control the choice of splitting or not. ‘minbucket’ Sets the minimum sum of weights in a terminal note, here set to 5% of the total weights in the training set. ‘minsplit’ Determines the needed minimum sum of weights in a node to consider a split, here set to 10% of the total weights in the training set.

The constructed tree yields the following breakpoints: 17, 26, 29, 32, 35, 38, 51, 55, 59, 63, 73, 95, which makes for 11 bins. This tree can also be graphed:



Clearly, the differences in expected claim frequency per bin can be seen in the boxplots below each bin in the tree. Policy holders younger than 26yo have a higher expected claim frequency compared to for example a 55yo policy holder. The binned variable is added to the original dataset under ‘agephGR’.

For severity …

(afbeelding boom)

# Frequency Analysis

Frequency analysis will be done using two models. On the one hand a Generalized Linear Model and on the other hand a Gradient Boosting Machine. Both will be compared whether they have sufficient predictive power, given their complexity. The final model will be chosen based on the higher accuracy on unseen data, while maintaining as simple as possible. A GLM is much simpler compared to a GBM, but is al lot less flexible.

## Generalized Linear Model

A GLM was chosen because it is one of the most interpretable models of all possible models, which is very important when explanability is an important factor as it is in the insurance industry. A simple model is easy to explain to the legislator and other stakeholders who may not be an expert in this field. As already mentioned, a GLM is not the most flexible model of the bunch. It requires linear relations and is preferably modelled with categorical independent variables. Because of this the spatial variable and the variable of the age of the policyholder were binned into two factor variables.

## Gradient Boosting Machine

## Conclusion