Territorial Ratemaking and Graph Theory

Marco De Virgilis

Disclaimer

The views and opinions expressed in this presentation are those of the authors and do not necessarily reflect the position of the organizations of which they are a part.

Agenda

- Theoretical Background
- Territorial Component
- Data
- Graph Analysis
- Model Comparisons
- Conclusion

Theoretical Background

Theoretical Background

- For many years, actuaries have recognized the importance of location as a major determinant of risk.
- Actuarial methodologies based on GIS (Geographical Information Systems) have been implemented and used predominately in the industry.
- Nowadays, almost all ratemaking methodologies for personal lines of automobile and homeowners insurance in the United States include a geographical component.

Ratemaking Model Structure

A standard pricing model can be summarized as:

$$Y = f(\beta * X, S(\cdot)) + \epsilon$$

Where:
- $\mathbf{Y}$ is the estimate of the expected loss cost.
- $\mathbf{f(\cdot)}$ is the function that defines the relation between the model components, e.g., additive, multiplicative.
- $\boldsymbol{\beta} * \mathbf{X}$ are the coefficients and the rating factors, e.g., Age, Gender, Car Brand, Credit Score.
- $\mathbf{S(\cdot)}$ is the Spatial component.
- $\boldsymbol{\epsilon}$ are the residuals or noise.

Territorial Component

Territorial Component

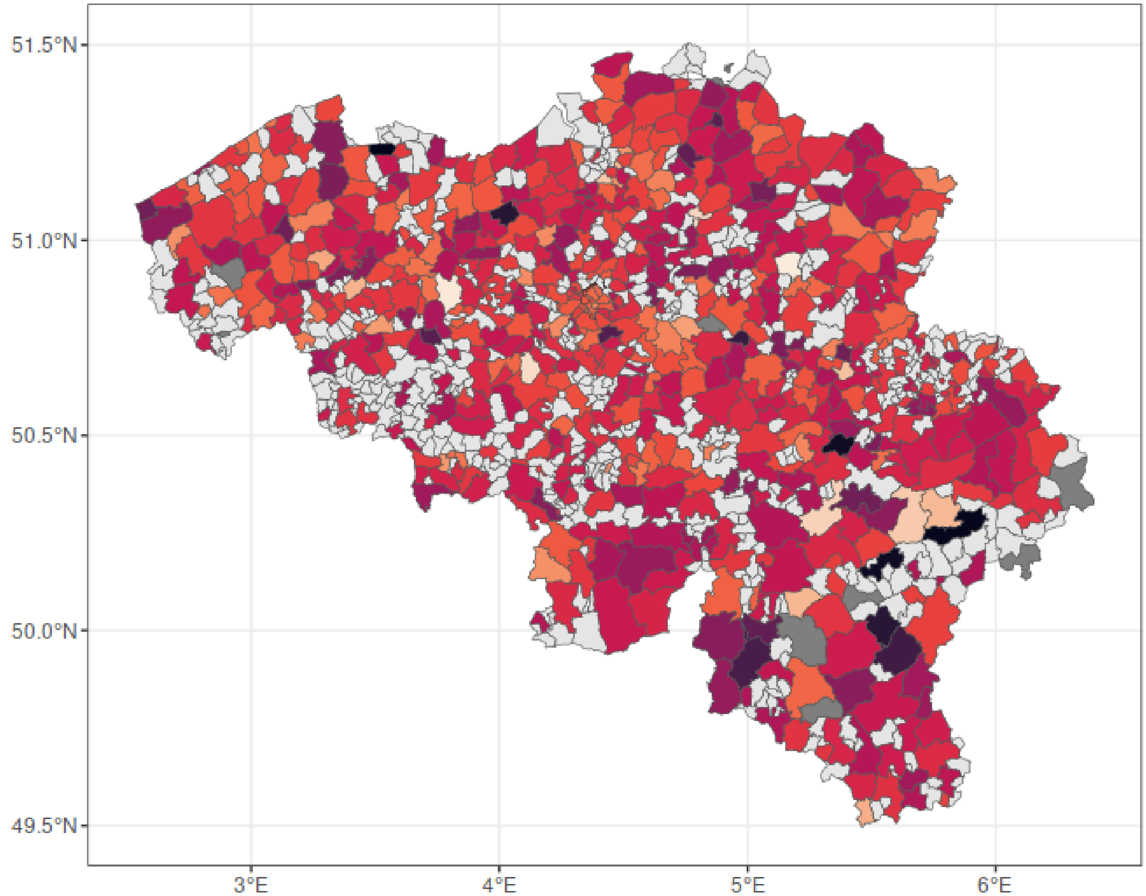
We will focus our attention on the $\mathbf{S}(\cdot)$ component of the model.
Two different approaches will be presented and evaluated; for simplicity and comparison purposes, we will keep the rating factors structure, i.e. $\beta * X$, identical between the two approaches.
The methodologies are:
- Spatial analysis entirely based only on geographical components, i.e. lat/long.
- Spatial analysis based on geographical components and enhanced with graph analysis.

Data

Data

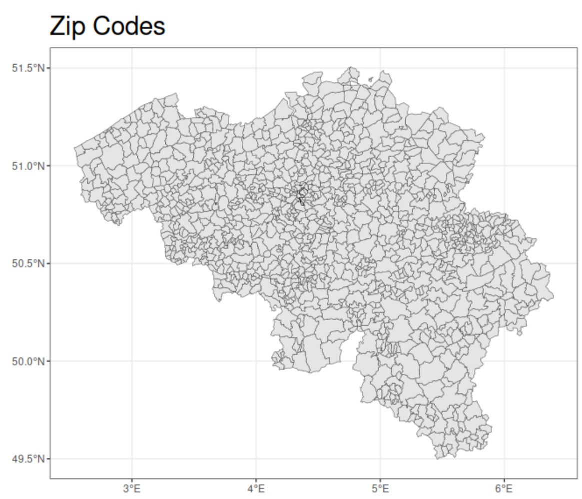
The data analyzed is a dataset of publicly available claims of Personal Auto Policies in Belgium.
The data contains standard rating factors, such as Age, Car Brand or Fuel Type alongside geographical information.

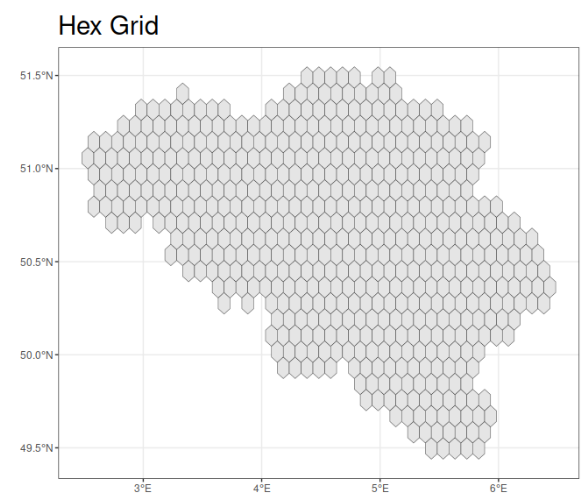
Loss Cost



Geographical Aggregation

The spatial analysis will be carried out at the zip code level.
This may not always be suitable or boundaries may not be available.
In such cases it will be necessary to produce a polygon grid to fill out the area; this will not affect the following analysis or the quality of the results.

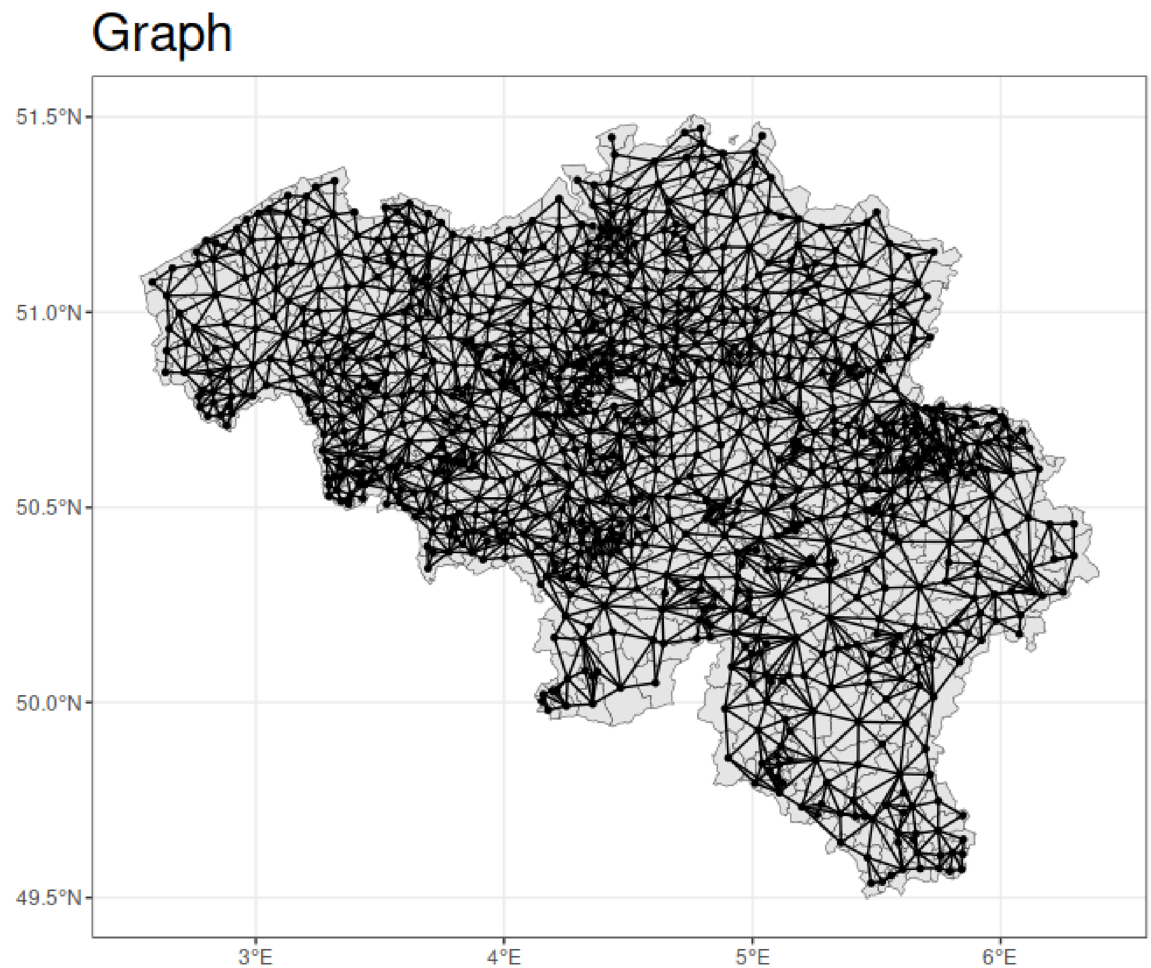




Graph Analysis

Graph Representation

The map of the territories can be interpreted as a graph with the following characteristics:
- The centroids of the individual territories will be the vertices.
- Each territory will be directly connected only to the adjacent ones.
- The length of the shared borders will be the weight of the edges.


Graph
51.5°N
51.0°N
50.5°N
50.0°N
49.5°N
3°E
4°E
5°E
6°E

Graph Characteristics

Having the data in this format allows to extend the features associated with the territories with elements stemming from graph theory. In particular, we will enhance territorial information with the centrality measures of the vertices. These are not the only information that are possible to extract, other examples may include: - Cliques
- Complete Subgraphs
- Coreness
- Edge Density
- …

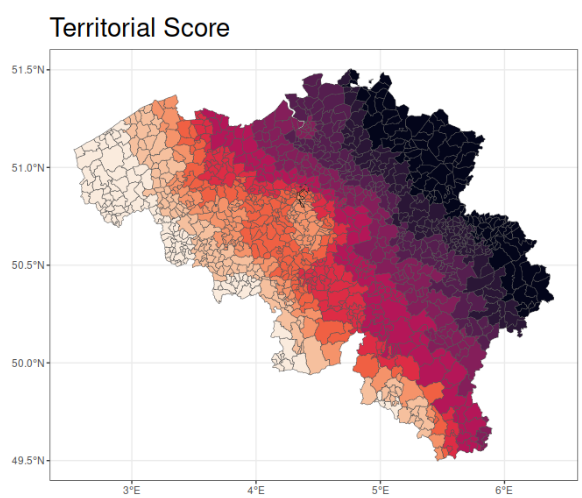
Centrality measures

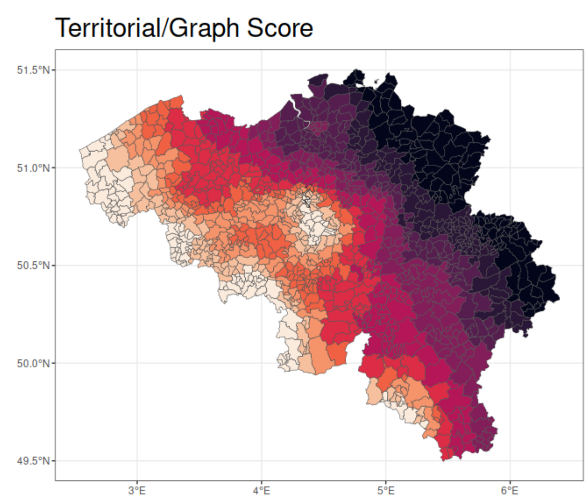
Centrality measures capture the importance of an edge in a graph. There are several types of centrality measures:
- **Degree centrality**: It measures the popularity of an edge in a network based on vertices.
- **Eigenvector centrality**: It measures the importance of an edge in a graph with respect to the importance of its neighbors.
- **Closeness centrality**: It tracks how close an edge is to another by measuring the distance between them.
- **Betweenness centrality**: It measures the importance of a node in a network based upon how many times it occurs in the shortest path between all pairs of nodes in a graph.

Model Comparison

Territorial Comparison

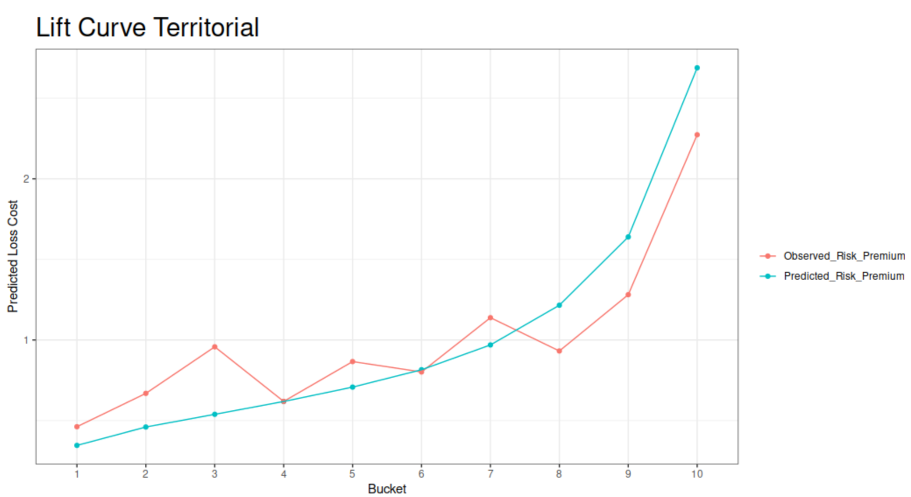
Territorial Clusters of similar risk have been computed employing spatial analysis techniques, namely Thin Plate Splines. Here the results of the geographical analysis:

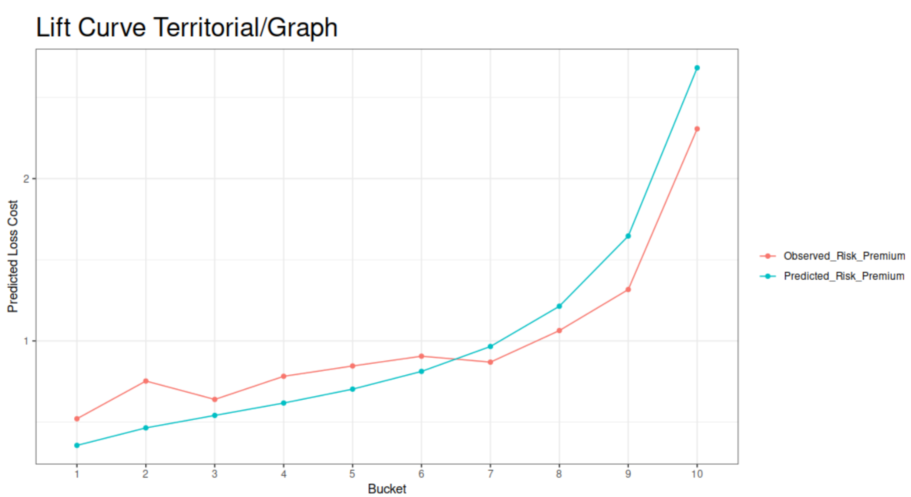




Overall Comparison

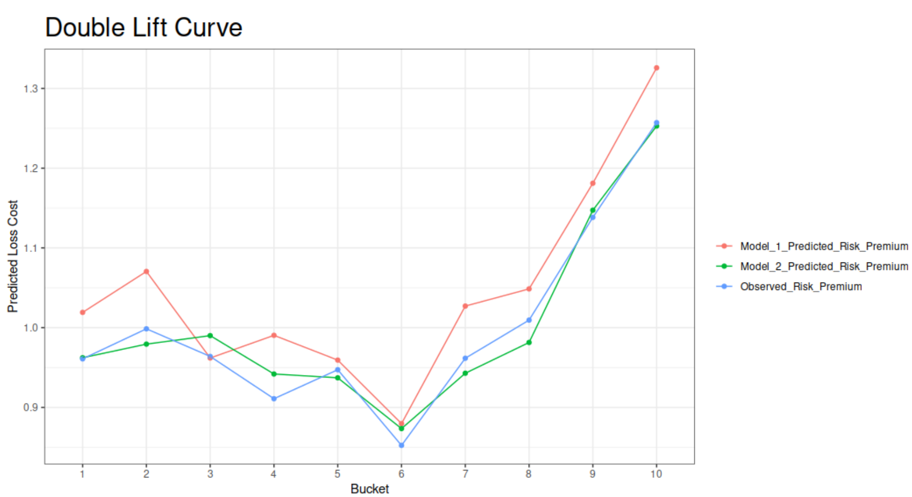
We then compared the predictive power of the full model in estimating the expected loss cost of the policyholder.
The model that includes the variables extracted from the graph is able to better capture the observed loss cost trend.

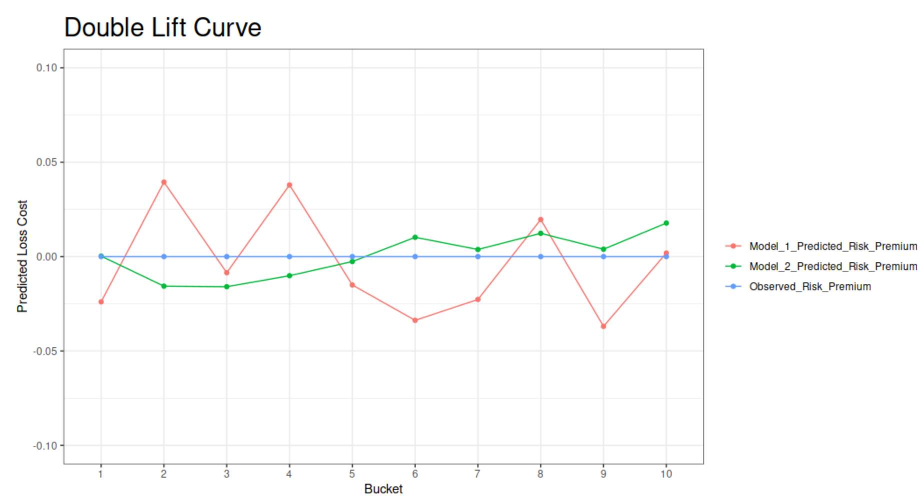




Overall Comparison

We then compared the predictive power of the full model in estimating the expected loss cost of the policyholder.
The model that includes the variables extracted from the graph is able to better capture the observed loss cost trend.





Conclusion

Conclusion

- The methodology, as presented, could certainly be improved and refined. Further studies could be focused on the implementation of different graph types and properties.
- Furthermore, it would be interesting to define similar structures that also take demographical characteristics into consideration in addition to the geographical ones, such as road types or traffic levels.

Contacts

Marco De Virgilis
devirgilis.marco@gmail.com