A Model Based Approach to Personal Auto Geographic Risk Classification

Jonathan Fesenmeyer and Matthew Berry
2021 RPM

Non-Geography Classifications
- Examples: Age, # of Operators, Vehicle Attributes
- Common variables and levels across states – can boost credibility with multistate analysis
- Typically, homogenous and credible loss experience without modification – can use widely accepted actuarial, modeling, or machine learning techniques

Geography Classifications
- Examples: County, Zip Code, Census Block
- Levels are unique to each state – can't boost credibility with multistate analysis
- Highly dimensional (~30k geographic Zip Codes in US)
- Majority of geography level experience isn't credible enough to be relied upon by itself without modification

Geography can't (or shouldn't) be analyzed the same way as other risk classifications

How do you represent an irregularly shaped 3D ellipsoid (Earth) on a 2D surface?
- You distort the lat/long coordinates by projecting it onto a 2D plane
- Every map you've ever used or created is "wrong", but many smart people have spent years creating techniques to make it "wrong" in the least impactful ways

Most GIS weirdness can be ignored if you know two things about your GIS data
- Know which projection you're using
  - EPSG 3857 is very common
- Know which CRS (Coordinate Reference System) you're using
  - WGS84 (EPSG 4326) is very common

Variety of software available to do GIS analysis
- R + a couple of R packages (sf and leaflet) is more than enough to do most GIS analysis

Fun Fact – For many years EPSG would not adopt the projection system developed/used by Microsoft and Google:
"We have reviewed the coordinate reference system used by Microsoft, Google, etc. and believe that it is technically flawed. We will not adopt the EPSG datum by including such inappropriate geodesy and cartography."
### Traditional Territory Technique

- **Smooth out Zip Level Noise by credibility smoothing with other nearby Zips’ Pure Premiums**
- **Cluster together Zips into Territories based on smoothed Zip Pure Premiums, often with a contiguity requirement**
- **Calculate each Territory’s Historical Pure Premium, often with additional credibility calculations**
- **Identify Non-Territory Factors from Pure Premium to convert to Indicated Territory Factor**

**Pros**
- Widely accepted and straightforward process

**Cons**
- Not contemporary with even relatively-modern techniques
- Attempts to capture territory effect exclusively through underlying experience
- Not multivariate so territory becomes a catch-all for deficiencies in the rating plan
- Difficult to validate accuracy beyond trust in the methodology and concepts

### Audience Question: Who created this visual?

**Actual Experience**
- **Signal**
- **Noise**

- **Non-Geographic**
- **Geographic**

### Contemporary Geography Technique

- Use multivariate approach (GLM) to get best estimate of true Geographic Signal
  - Rating plan variables act as controls
  - No longer a catch-all for the rest of the rating plan
  - Robust holdout/validation framework

- Use Explanatory and Non-Explanatory methods to capture Geographic signal
  - Explanatory variables (geovariables) act as loss predictors that describe the demographic, weather, and environmental features of a given geographic area
    - E.g. population density, median commute times, annual rainfall, etc
  - Non-Explanatory variables (clusters) act as loss predictors to capture the still important pure geographic effect that isn’t explained by other variables
The idea for a better mousetrap

- Explanatory effects (geovariables) generally capture more signal, more efficiently, than non-explanatory effects
  - Non-explanatory effects are more prone to overfitting by their nature
- Takeaway - Capture as much signal through geovariables as possible

Lightbulb
- Insureds don’t just drive in their Zip!
- They drive in nearby Zips and get exposed to those drivers and conditions (and vice versa)!

- Geovariables can be made even more powerful
  - Spatially smooth each Zip’s geovariable with the geovariables from nearby Zips (e.g. within a certain distance)
  - Goldilocks problem
    - Too little smoothing – noisy and don’t get predictive lift
    - Too much smoothing – compresses spread and masks signal
  - Get it just right – easier to analyze and capture signal
- Need to understand the spatial relationship of every Zip Code…... so.......

Building said better mousetrap

How to calculate distances between every pair of Zips (polygons)
- Centroid-to-Centroid Distance
  - Variation in Zip sizes and shapes can severely bias this measure
- Vertex-to-Vertex Distance
  - Too much information lost
- Sampled Point-to-Sampled Point Distance
  - Sample points around Polygon (Zip) edges and calculate distance between all points
  - Take min or max for each pair of Zips to approximate to the min or max distance
  - Computationally intensive but creative coding can help

Audience Question:
Given many possible predictor choices.... How do you find the best ones?
Geovariables here, there, everywhere

- Problem — there can be hundreds of geovariables (plus rating plan controls) to choose from
- How do we find the most predictive ones?
- Lots of possible analyses to try to answer this
- All have pros/cons
- In our testing, some of the more popular approaches don’t hold up to this specific problem

Exploratory Lasso Variable Selection Technique

- Lasso can regularize coefficients to absolute zero (variable selection)
- Provides a 90% starting point for the GLM
- Not perfect or full automation, but a useful tool
- Found that by using this technique after a traditional GLM build process —
  this lasso identified almost all the variables that were manually identified
- The lasso identified several valuable variables that were overlooked or ignored during the manual GLM build
- The lasso identified few variables that upon further examination didn’t seem worthwhile or were not highly predictive

Completening the Explanatory
Starting the Non-Explanatory Journey...

- So, you’ve built a great GLM using your fancy smoothed geovariables you discovered from your new Lasso Exploratory analysis...
- Explanatory effects capture a lot of the signal, but not all of it
- Non-Explanatory effects (pure geography) not explained by geovariables can still provide significant value
- But we’re now firmly back in the land of needing a specialized approach given the uniqueness of geography

- Early steps in the Non-Explanatory Journey
  - Score the GLM, aggregate actual and predicted to Zip level to get Residual Pure Premium
  - Residuals are noisy and can have credibility concerns
  - Use knowledge of spatial relationship between Zips and credibility concepts to check usefulness of standardized residuals
  - Goal is to smooth the least amount necessary to avoid overfitting in later steps
• Smoothed Zip Residuals are still too noisy to use directly
  - Need to aggregate them into a smaller number of more credible groupings (aka dimensionality reduction)
• K-Medoids Clustering - Very similar to widely used and accepted K-Means method (but is a little more resilient to overfitting on outliers)
  - Even still, clustering can overfit extremely easily if you aren't careful
  - We need a way to protect the clustering algorithm from itself
    - Knowledge of spatial relationships between Zips to the rescue
    - Incentivize the algorithm to group together nearby Zips vs far away Zips to prevent overfitting
    - Be able to control how much of this incentive is given
      - Don’t want to overfit, but don’t want to underfit either
---

### Lightbulb
- Knowledge of spatial relationships between Zips to the rescue
- Incentivize the algorithm to group together nearby Zips vs far away Zips to prevent overfitting
- Be able to control how much of this incentive is given
  - Don’t want to overfit, but don’t want to underfit either

### Effect of PP Weight (w) Tuning Parameter on Clustering

- Pure Geography Clustering
- Similarity Measure = Minimum_Zip_Mile_Distance

- Pure Loss Residual Clustering
- Similarity Measure = Loss_Residual_Distance

- General Mixture Clustering
  - Similarity Measure = $w \times \text{Loss}_\text{Res}_\text{Dist} + (1-w) \times \text{Min}_\text{Zip}_\text{Mile}_\text{Dist}$

- Lightbulb
  - Knowledge of spatial relationships between Zips to the rescue
  - Incentivize the algorithm to group together nearby Zips vs far away Zips to prevent overfitting
  - Be able to control how much of this incentive is given
    - Don’t want to overfit, but don’t want to underfit either

---

### Non-Explanatory Journey: Extended Edition 4 Disc Set

- Two selections for this Non-Explanatory approach:
  - Number of clusters (N)
  - Weight given to residual vs spatial distance (W)
- How to choose? One approach given enough compute capacity and coding capabilities
  - Create many clustering solutions for different combinations of W and N
  - Evaluate goodness of fit of each solution
  - Choose the clustering solution with the highest goodness of fit

- The value of Non-Explanatory effect
  - It depends (sorry)
    - If clustering adds enough value, add the solution to your model
    - If clustering doesn’t add much value, remove it
    - Clustering decisions are very specific to your data and model structure
      - Some coverages may benefit more from clustering than others

---

2/21/2021
One little trick in model scoring is all that’s needed to convert models to factors.

Results

- More accurate factors
  - Can be proven through common model validation techniques
  - Increased spread in factors
- More granularity, even in less populated areas
  - Able to move past the “all other” bucketing that traditional approaches can rely on
- Geovariables result in unique factors for each Zip