A Model Based Approach to Personal Auto Geographic Risk Classification

Jonathan Fesenmeyer and Matthew Berry

2021 RPM

Modeling & Major Initiatives

Pricing Analytics & Actuarial Services

1

Why Geography Unique in Risk Classification Systems

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
  - We'll focus on Zip Code throughout
- 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

Modeling & Major Initiatives

Pricing Analytics & Actuarial Services 2 Internal Information

2

~~Acronym Soup~~ GIS Intro (tl;dr - maps are complicated)

- 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

EPSG (European Petroleum Survey Group) maintains these projection techniques

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 devalue the EPSG dataset by including such inappropriate geodesy and cartography."*

Modeling & Major Initiatives

Pricing Analytics & Actuarial Services 3 Internal Information

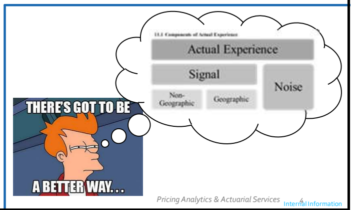
3

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 ⇨ **Backout Non-Territory Factors** from Pure Premium to convert to Indicated Territory Factor

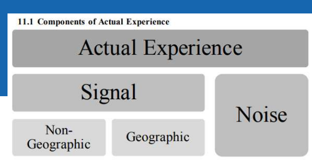
Pros
- Widely accepted and straightforward process

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



4

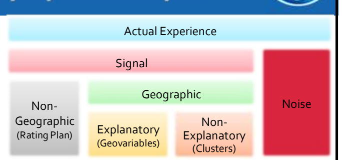
Audience Question:
Who created this visual?



5

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



Modified from Werner & Modlin's Basic Ratemaking

- 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

6

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)!

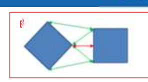
- Geovariables can be made *even* more powerful
  - Spatially smooth each Zip's geovariable with the geovariables from nearby Zips (e.g. within a certain distance)
  - Goldilocks problem
    - Too little smoothing – noisy and don't get predictive lift
    - Too much smoothing – compresses spread and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

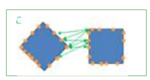
7

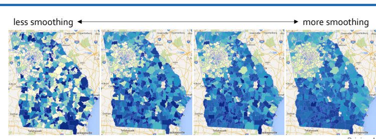
Building said better mousetrap

How to calculate distances between every pair of Zips (polygons)

A. Centroid-to-Centroid Distance
   - Variation in Zip sizes and shapes can severely bias this measure
B. Vertex-to-Vertex Distance
   - Too much information lost
C. 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







8

Audience Question:
Given many possible predictor choices....
How do you find the best ones?

9

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 pro's/con's
  - 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)

Ridge

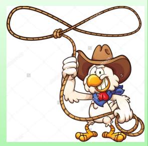
$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

Elastic Net (GLMnet)

$$\mathrm{RSS} + \lambda\left(\alpha \sum_{j=1}^{p} \beta_j^2 + (1-\alpha) \sum_{j=1}^{p} |\beta_j|\right)$$

Lasso

$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$



10

Exploratory Lasso Variable Selection Technique

Setup
- Setup Lasso GLM
- Pass all candidate variables (and transformations of candidate variables) as predictors into the model
- Run across a wide range of Lambdas (which controls the regularization strength)

Evaluation
- Examine prevalence of non-zero coefficients across Lambdas, by variable
  - More powerful variables tend to be non-zero even at higher Lambdas
  - Less powerful variables tend to get zero'd out
- Look for patterns of which variables (or levels) tend to be consistently non-zero as likely candidates for GLM
- Lasso handles multiple versions of a similar variable well by tending to allocate the effect entirely to 1 or 2 variables (plus general stability in different runs)

coefficients by lambda	0.000	0.008	0.016	0.024	0.032	0.040	0.080	0.110	0.140	0.170	0.200
g_Pct_Built_After1940_R_5mi_floor75	-0.009	0.006	0.006	0.007	0.006	0.005	0.003	0.002	0.002	0.002	0.001
g_Pct_Built_After1950_R_5mi_floor70	0.01	0	0	0	0	0	0	0	0	0	0
g_Pct_Built_After1960_R_5mi_floor65	0.022	0	0	0	0	0	0	0	0	0	0
g_Pct_Built_After1970_R_5mi_floor55	-0.049	0	0	0	0	0	0	0	0	0	0
g_Pct_Built_After1980_R_5mi_floor45	0.054	0.007	0.002	1E-04	0	0	0	0	0	0	0
g_Pct_Built_After1990_R_5mi_floor30	-0.03	0	0	0	0	0	0	0	0	0	0
g_Pct_Built_After2000_R_5mi_floor_15	-0.003	-0.005	-0.002	0	0	0	0	0	0	0	0
g_Pct_Built_After2005_R_5mi_5to15	0.009	0	0	0	0	0	0	0	0	0	0
g_Pct_Built_Before1940_R_5mi	0.042	0.021	0.012	0.007	0.005	0.003	0	0	0	0	0
g_Pct_Commute_Leave_Within_Rush_Hour_6am_R_5	-0.174	0	0	0	0	0	0	0	0	0	0
g_Pct_Commute_Leave_Within_Rush_Hour_7am_R_5	-0.576	0	0.144	0.087	0.042	0.02	0	0	0	0	0
g_Pct_Commute_Travel_Over90min_R_5mi	-0.776	0	-0.269	-0.107	0	0	0	0	0	0	0
g_Pct_Commute_Travel_Under10min_R_5mi_cap20	3.043	0	0	0	0	0	0	0	0	0	0
g_Pct_Commute_Travel_Under15min_R_5mi_cap40	-2.18	-0.906	-0.657	-0.54	-0.47	-0.416	-0.183	-0.135	-0.147	-0.144	-0.137
g_Pct_Commute_Travel_Under20min_R_5mi_cap60	0.338	0	0	0	0	0	0	0	0	0	0
g_Pct_Commute_Travel_Under25min_R_5mi	0.37	0	0	0	0	0	0	0	0	0	0
g_Pct_Commute_Travel_Under30min_R_5mi	-1.404	0	0	0	0	0	0	0	0	0	0
g_Pct_Commute_Travel_Under35min_R_5mi	0.77	0	0	0	0	0	0	0	0	0	0
g_Pct_Commute_Travel_Under40min_R_5mi	2.404	0.15	0	0	0	0	0	0	0	0	0
g_Pct_Commute_Travel_Under45min_R_5mi	1.392	0	0	0	0	0	0	0	0	0	0
g_Pct_Commute_Travel_Under5min_R_5mi_cap6	-5.158	-2.113	-2.643	-2.747	-2.958	-3.095	-3.76	-3.769	-3.547	-3.437	-3.398
g_Pct_Commute_Travel_Under60min_R_5mi	-5.755	0	0	0	0	0	0	0	0	0	0
g_Pct_Commute_Travel_Under90min_R_5mi	0.776	0	0.269	0.107	0	0	0	0	0	0	0

- Found that by using this technique after a traditional GLM build process – the 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 a few variables that upon further examination didn't seem worthwhile (false positives)

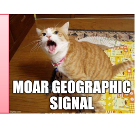
Strongly recommend Exploratory Lasso to aid in the building of traditional GLMs
- Not perfect or full automation, but a useful tool
- Provides a 90% starting point for the GLM
- Dramatically quicker path to better insights

11

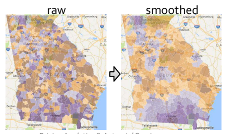
Completing the Explanatory
Starting the Non-Explanatory Journey…

- 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 create Smoothed Residuals
    - Goal is to smooth the least amount necessary to avoid overfitting in later steps



12

Continuing the Non-Explanatory Journey

- 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

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

Pure Loss Residual Clustering
Similarity Measure = Loss_Residual_Distance

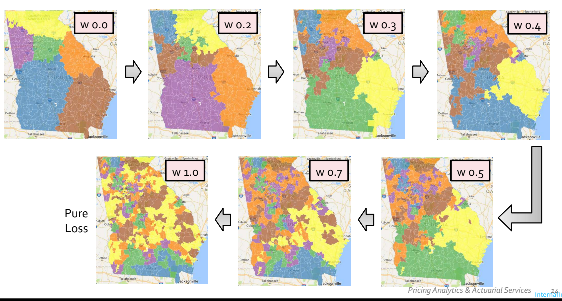
Pure Geography Clustering
Similarity Measure = Minimum_Zip_Mile_Distance

General Mixture Clustering
Similarity Measure = w*Loss_Res_Dist + (1-w)*Min_Zip_Mile_Dist
where $0 \leq w \leq 1$

Modeling & Major Initiatives

13

Effect of PP Weight (w) Tuning Parameter on Clustering

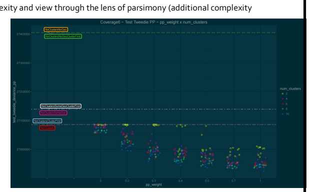


Modeling & Major Initiatives

14

Non-Explanatory Journey: Extended Edition 4 Disc Set

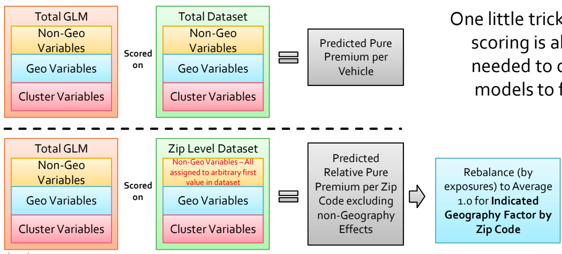
- Two selections for this Non-Explanatory approach:
  - Number of clusters (N)
  - Weight given to Residual vs Spatial Distance (W)
  - How to choose? One approach given enough compute capacity and coding capabilities
    - Create many clustering solutions for different combinations of N and W
    - Fit a version of the GLM for each clustering solution, with the cluster solution just being an additional variable in the model
    - Evaluate goodness of fit diagnostics to zone in on N and W values
      - Beneficial to think of larger N and W values as increased model complexity and view through the lens of parsimony (additional complexity needs to really provide benefit)
- The value of Non-Explanatory effect
  - It depends (sorry!)
  - If clustering adds enough value, add the solution to your model
    - Sometimes the best solution is to not use any of the clusters and stick with geovariables (especially if you did a great job on that part of it) if clustering doesn't add enough value or makes things worse
  - Clustering decisions are very specific to your data and model structure
    - Some coverages seem to get a lot more value than others (Comprehensive notably benefits quite a bit from clustering)



Modeling & Major Initiatives

15

Converting Models to Geography Factors

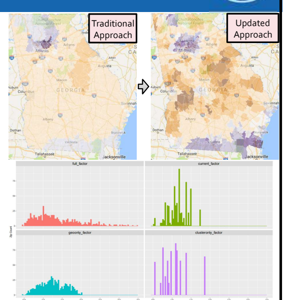


One little trick in model scoring is all that's needed to convert models to factors

16

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



17