Territorial risk classification using spatially dependent frequency-severity models

Peng Shi

University of Wisconsin - Madison

March 15, 2016

Outline

- Non-life insurance often classifies risks geographically
- It is important for insurance operations
  - Marketing, underwriting, ratemaking ...
- Risk classification could be formulated in a regression setup
  - Common practice uses GLMs, e.g. frequency-severity and pure premium models
  - Claims model is built using micro-level (policy or claim) data
- Our goal: to build a claims model to create territory-level risk scores
  - Use aggregate claims data and use the two-part framework
  - Account for the spatially correlation and the association between frequency and severity
  - The score is used to supplement the claims modeling with micro-level data
  - We demonstrate applications in prediction and market segamentation

Two-part Specification

- Frequency model
  - We model the number of policyholders in region i that incur at least one claims, denoted by Y_i^f
  - Assume a binomial distribution

$$Y_i^f \sim Bin(E_i, p_i)$$
$$\log\left(\frac{p_i}{1-p_i}\right) = x_i^f \boldsymbol{\beta}_f + \phi_i^f$$

- Severity model
  - we model the average amount of payment per policy in region i given occurrence of claims, denoted by Y_i^s
  - Assume a log-normal distribution

$$Y_i^s \sim LN(\mu_i, \tau^{-1})$$
$$\mu_i = x_i^s \boldsymbol{\beta}_s + \phi_i^s$$

Examples of Risk Score

Table : Examples of risk scores

Example	Score Function
1	$g(\theta_i^f, \theta_i^s) = p_i$
2	$g(\theta_i^f, \theta_i^s) = p_i/(1 - p_i)$
3	$g(\theta_i^f, \theta_i^s) = \exp(\mu_i + \frac{1}{2}\tau^{-1})$
4	$g(\theta_i^f, \theta_i^s) = p_i \exp(\mu_i + \frac{1}{2}\tau^{-1})$
5	$g(\theta_i^f, \theta_i^s) = \sqrt{p_i}/(\exp(\tau^{-1}) - p_i)$

Spatial Specification: Univariate CAR

- Conditional distribution

$$\phi_i|\phi_{-i} \sim N\left(\frac{\gamma}{m_i}\sum_{i\sim j}\phi_j, \frac{1}{\lambda m_i}\right), \; i=1,\ldots,n$$

  - γ spatial dependence, λ spatial dispersion
  - m_i denote number of neighbors for region i

- Joint distribution

$$\boldsymbol{\phi} \sim N_n(0, [\lambda(D-\gamma W)]^{-1})$$

  - $D = \text{diag}(m_1, \ldots, m_n)$
  - W is the adjacency matrix, $w_{ii}=0$, $w_{ii'}=1$ if $i \sim i'$

Spatial Specification: Bivariate CAR

- Let $\boldsymbol{\phi} = (\boldsymbol{\phi}'_f, \boldsymbol{\phi}'_s)'$ and $\mathbf{v} = (\mathbf{v}'_f, \mathbf{v}'_s)'$. Based on linear model of co-regionalization (LMC)

$$\boldsymbol{\phi} = (B \otimes I_{n\times n})\mathbf{v}$$

  - $\mathbf{v}_f$ and $\mathbf{v}_s$ are two independent latent spatial processes
  - B is non-singular

- Consider two cases
  - Separable model: $\mathbf{v}_f$ and $\mathbf{v}_s$ are identical
  - Inseparable model: $\mathbf{v}_f$ and $\mathbf{v}_s$ are not identical

Spatial Specification: Bivariate CAR

$$\boldsymbol{\phi} \sim N_{2n}(0, \Omega)$$

- $\mathbf{v}_f \sim N_n(0, (D-\gamma W)^{-1})$, $\mathbf{v}_s \sim N_n(0, (D-\gamma W)^{-1})$
  - $\Omega = [(D-\gamma W) \otimes \Lambda]^{-1}$
  - Identifiable up to $\Lambda^{-1} = BB'$
- $\mathbf{v}_f \sim N_n(0, (D-\gamma_f W)^{-1})$, $\mathbf{v}_s \sim N_n(0, (D-\gamma_s W)^{-1})$
  - $\Omega = (B \otimes I_{n\times n})(I_{2\times 2}D - \Gamma \otimes W)^{-1}(B \otimes I_{n\times n})'$
  - $\Gamma = \mathrm{diag}(\gamma_f, \gamma_s)$
- Define $\Sigma = BB'$, use B as upper triangular Cholesky decomposition of Σ

Data

- Personal automobile insurance data from the Commonwealth Automobile Reinsurers (CAR) in Massachusetts
  - The data represent experience from several insurance carriers
  - The dataset contains claims records about two million policyholders in year 2006
  - Claims data on two mandatory coverage
    - Liability and PIP
    - We look at combined coverage
  - Limited information on predictors
    - Rating group: policyholder characteristics
    - Territory group: defined by garage town (351 towns in Massachusetts)
- Info on vehicle characteristics is supplemented by ISO
  - Vehicle age, car type, other features...

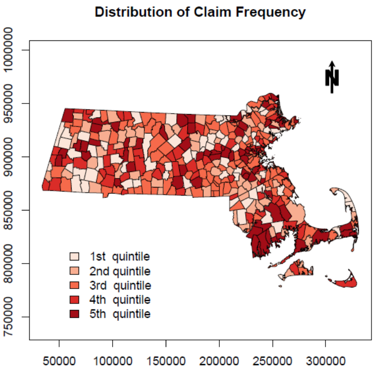
Sampling

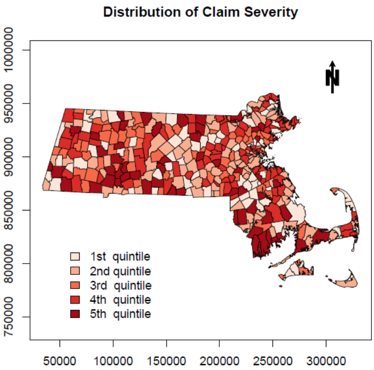
- Use 80% of data to build the model
  - Stratified sampling
  - Aggregate claims data at town level
- Use the rest 20% of data for application at micro-level observations
  - Use 75% to build model at policy/claim level
  - Use 25% for out-of-sample validation

Table : Descriptive statistics of outcomes and covariates

Variable	Description	mean	std	min	max
Response variable					
freq	Frequency of at least one claims	3.66	1.10	0.00	7.80
size	Average size of payments	3250.08	832.45	678.78	7291.09
Covariates					
young	Percentage of young driver	9.93	1.80	3.13	22.22
senior	Percentage of senior driver	15.01	5.10	0.00	39.03
vehage	Average vehicle age	5.40	0.27	4.56	6.67
lux	Percentage of luxury car	4.69	2.83	0.00	20.95
van	Percentage of van	7.85	1.63	3.80	22.22
pickup	Percentage of pickup truck	14.27	6.23	1.09	33.33
utility	Percentage of utility vehicle	24.76	4.42	11.11	58.14
awd	Percentage of vehicle with all wheel drive	41.65	9.28	26.93	77.08

Summary Statistics





Estimation

	Independent Aspatial Model		Bivariate Spatial Model	
	Estimate	95% Credible Interval	Estimate	95% Credible Interval
Frequency Model				
Intercept	-3.532	(-3.727, -3.337)	-3.211	(-3.713, -2.691)
young	-1.837	(-2.351, -1.315)	-0.862	(-2.088, 0.370)
senior	-1.670	(-1.850, -1.485)	-1.536	(-1.919, -1.144)
vehage	0.293	(0.264, 0.322)	0.227	(0.148, 0.304)
lux	2.429	(1.913, 2.931)	1.809	(0.740, 2.845)
van	-3.824	(-4.285, -3.347)	-4.144	(-5.372, -2.925)
pickup	-1.792	(-2.030, -1.549)	-1.818	(-2.361, -1.293)
utility	3.064	(2.780, 3.351)	1.943	(1.362, 2.539)
awd	-3.038	(-3.262, -2.813)	-2.387	(-2.788, -1.979)
Severity Model				
Intercept	8.142	(7.000, 9.327)	7.634	(7.163, 8.158)
young	-2.051	(-4.173, 0.103)	-1.719	(-3.734, 0.293)
senior	-0.670	(-1.395, 0.065)	-0.633	(-1.342, 0.059)
vehage	0.047	(-0.131, 0.215)	0.111	(0.014, 0.201)
lux	2.933	(0.800, 5.005)	3.333	(1.546, 5.187)
van	1.163	(-1.504, 3.806)	1.827	(-0.498, 4.126)
pickup	1.470	(0.442, 2.491)	1.544	(0.572, 2.560)
utility	1.767	(0.758, 2.798)	1.881	(0.857, 2.844)
awd	-2.212	(-2.813, -1.582)	-2.170	(-2.819, -1.444)
Dispersion	13.720	(11.730, 15.940)	17.420	(13.930, 22.710)
Dependence Model				
α_f			0.202	(0.009, 0.518)
α_s			0.463	(0.025, 0.934)
σ_f			16.930	(13.190, 21.410)
σ_s			10.640	(0.363, 35.220)
ρ			0.833	(0.414, 0.998)

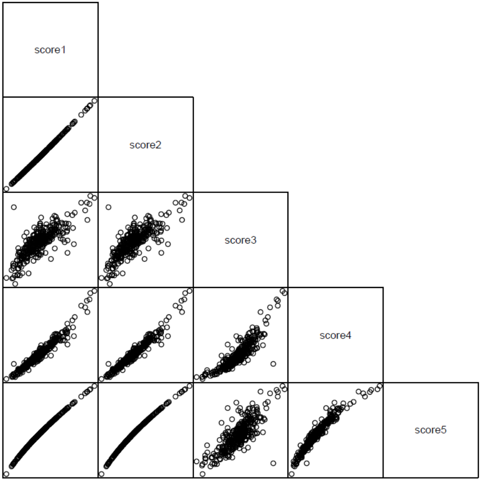
Model Comparison

Table : Goodness-of-fit statistics for alternative models

Model	Description	DIC
1	Independent aspatial model	9,661
2	Independent spatial model	8,463
3	Intrinsic bivariate spatial model	8,361
4	Separable bivariate spatial model	8,323
5	Inseparable bivariate spatial model	8,262

Summary of Scores

- Correlation between different scores
- Scores are calculated using posterior mean of parameters



Prediction

- Use the score to supplement model building with micro-level data
- We compare out-of-sample prediction with and without score for different model specifications
  - Two-part model
    - Policy level: Logit + LN
    - Claim level: Poisson + Gamma
  - Pure premium model: Tweedie GLM
  - Use driver and car characteristics as predictors
- Out-of-sample prediction is evaluated using Gini statistics

Table : Out-of-sample validation for frequency-severity models

	Gini Correlation	Simple Gini
FrequencyModel		
Logit: policy info only	3.501	4.252
Logit: policy info + town	6.000	7.288
Logit: policy info + score.freq	6.517	7.915
Poisson: policy info only	3.526	4.593
Poisson: policy info + town	6.108	7.957
Poisson: policy info + score.freq	6.571	8.560
SeverityModel		
LN: policy info only	2.668	100.073
LN: policy info + town	4.703	176.489
LN: policy info + score.sev	5.235	196.398
Gamma: policy info only	4.190	149.798
Gamma: policy info + town	5.554	198.644
Gamma: policy info + score.sev	6.128	219.134

Table : Out-of-sample validation for pure premium models

	Gini Correlation	Simple Gini
Tweedie GLM: policy info only	2.467	22.788
Tweedie GLM: policy info + town	4.015	37.095
Tweedie GLM: policy info + score.pp (offset)	4.287	39.611
Tweedie GLM: policy info + score.pp	4.287	39.613
Tweedie GLM: policy info + score.freq + score.sev	4.289	39.626
Two-part: Logit + LN	4.184	38.662
Two-part: Poisson + Gamma	4.288	39.618

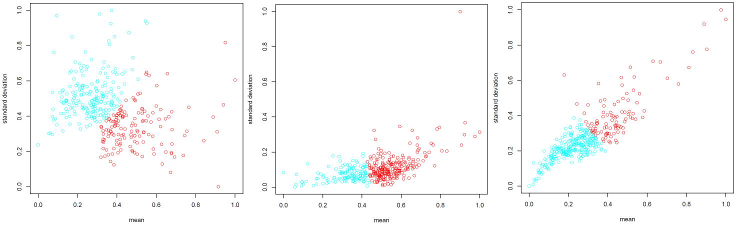
Market Segmentation

- Clustering customers for marketing purposes
- The model output can be used in clustering in a hierarchical manner

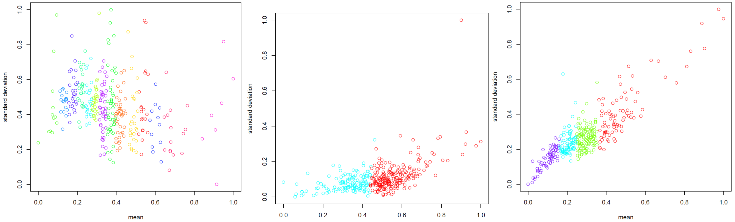
- Using K-medoids for clustering (Partitioning Around Medoids)
- Optimal number of clustering is based on Silhouette coefficient
- Based on mean/variance (Euclidean distance) and the entire distribution of the score (Jensen Shannon distance)

Market Segmentation

- Mean/variance: frequency, severity, and pure premium

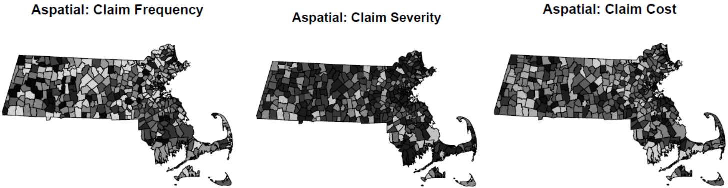


- Distribution: frequency, severity, and pure premium

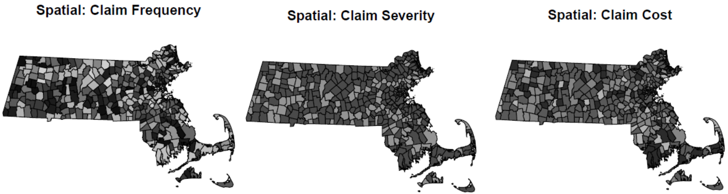


Market Segmentation

- Aspatial model clustering



- Spatial model clustering



Conclusion

- We built a frequency-severity model using region-level aggregated claims data
- The model took into account the correlation across space as well as between frequency and severity components
- We demonstrated some applications in prediction and market segmentation based on the main output of the model - territory risk score

Thank you for your kind attention.

Learn more about my research at:
`https://sites.google.com/a/wisc.edu/peng-shi/`