TERRITORIAL RISK CLASSIFICATION USING SPATIALLY DEPENDENT FREQUENCY-SEVERITY MODELS

BY

PENG SHI AND KUN SHI

ABSTRACT

In non-life insurance, territory-based risk classification is useful for various insurance operations including marketing, underwriting, ratemaking, etc. This paper proposes a spatially dependent frequency-severity modeling framework to produce territorial risk scores. The framework applies to the aggregated insurance claims where the frequency and severity components examine the occurrence rate and average size of insurance claims in each geographic unit, respectively. We employ the bivariate conditional autoregressive models to accommodate the spatial dependency in the frequency and severity components, as well as the cross-sectional association between the two components. Using a town-level claims data of automobile insurance in Massachusetts, we demonstrate applications of the model output—territorial risk scores—in ratemaking and market segmentation.

KEYWORDS

Bivariate CAR, clustering, insurance claims, spatial dependence.

1. INTRODUCTION

In non-life insurance, individual risks are often correlated spatially. For instance, claims in homeowner insurance often display geographic patterns due to weather-related hazards. Similar arguments also apply to automobile insurance where the frequency and severity of car accidents are highly influenced by factors such as traffic density and road conditions. In health insurance, care expenditures present spatial correlation because local health status and health care utilization are affected by the socio-economic profile, population density and aging. Modeling the spatial dependency and understanding its effects on an insurer's portfolio help to balance its risk pool geographically in choosing what risks to insure against and what coverage to offer, and help to assess the geographic relation and build the information into pricing of its insurance contracts.

Astin Bulletin 47(2), 437-465. doi: 10.1017/asb.2017.7 © 2017 by Astin Bulletin. All rights reserved.

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

The research question that we are interested in is how an insurer could leverage industry-level information into its own operations such as marketing, underwriting or ratemaking. One can think that an insurer can easily obtain the industry data from sources such as professional associations, government agencies, bureau of statistics, etc. Motivated by the territorial risk classification in non-life insurance, our goal is to produce a risk score for each geographical region. In this paper, we develop the score using macro-level data collected over geographical regions and we demonstrate the applications using micro-level data observed for individual policyholders. To build the score, this paper proposes a spatially dependent frequency-severity framework for modeling insurance claims in non-life insurance. Following the standard two-part approach, we decouple insurance claims into the frequency and severity components. To capture the spatial heterogeneity in each component, we consider the bivariate conditional autoregressive (CAR) models based on linear models of co-regionalization. The proposed model easily accommodates the spatial dependency in the frequency and severity components, as well as the cross-sectional association between the two components.

Our work relates to two strands of literature in insurance claims modeling. The first is spatial modeling of frequency and severity. Despite the rich development of spatial statistics in disciplines such as biology and epidemiology, the literature on spatial modeling of insurance claims is still sparse, presumably due to some unique features of the insurance data. The first effort in this line of study is Dimakos and Di Rattalma (2002). They employed the spatial Poisson and spatial gamma models to examine the claims from car theft in Norway, and unfortunately the proposed spatial structure was not supported by the data. Gschlößl and Czado (2007) extended the approach in Dimakos and Di Rattalma (2002) by using the observed claim count as a predictor for the average claim severity, and thus allowing dependence between them. As in Dimakos and Di Rattalma (2002), they also considered the CAR model for the spatial structure, and used independent specifications in the frequency and severity components. Another example is Czado *et al.* (2014), where the authors investigated the model selection issue for spatial count regressions in the context of health insurance. We refer to Brechmann and Czado (2014) for an introduction of spatial modeling in actuarial applications.

The second is dependent modeling of frequency and severity. Popular methods for pricing non-life insurance contracts, including the frequency-severity or two-part model and the Tweedie compound Poisson or pure premium model, often assume an independent relationship between the frequency and severity of claims. There are a few efforts in the current literature toward relaxing this independence assumption. As noted above, Gschlößl and Czado (2007) considered the regression of claim size on claim count in automobile insurance. This conditional probability approach has also been used in modeling health care expenditures, for example, see Frees *et al.* (2011) and Erhardt and Czado (2012). Czado *et al.* (2012) and Krämer *et al.* (2013) employed parametric copulas to jointly model the number and the average size of claims using aggregated

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

car insurance data. Shi *et al.* (2015) compared the predictive performance of both approaches using simulation and micro-level claims data in automobile insurance.

What sets our work apart from existing literature is the joint modeling of the spatial random effects in the frequency and severity components. Instead of assuming independent CARs, we explore bivariate CARs that allow for flexible spatial structure within the frequency and severity models, in the meanwhile, inducing the cross-sectional correlation between them. We also contribute to the literature by demonstrating novel applications of the model output—a risk score for each areal unit—using a unique dataset of automobile insurance in Massachusetts.

The rest of the article is structured as follows: Section 2 introduces the Massachusetts automobile insurance claims database and explores the data characteristics that motivate the statistical model. Section 3 details the spatially dependent frequency-severity model and the Bayesian inference procedure. Section 4 summarizes the estimation and model selection results based on the aggregated claims data on areal units. Section 5 presents two applications of the model output, ratemaking and market segmentation. Some technical details and supplementary analysis are provided in appendices to make the paper self-contained.

2. Data

We use a database of automobile insurance obtained from the Commonwealth Automobile Reinsurers in the state of Massachusetts in the United States. The Commonwealth Automobile Reinsurers is an industry-operated residual market and statistical agent for motor vehicle insurance in the Commonwealth of Massachusetts. As a statistical agent, the Commonwealth Automobile Reinsurers collects, edits and processes premium and claim data of private passenger and commercial automobile insurance in Massachusetts.

We analyze a claim dataset of personal automobile insurance. The dataset contains claims records about two million policyholders corresponding to the minimum required coverage in the year 2006. We focus on policyholders with a full year exposure to simplify the measurement of frequency and severity. For each individual policyholder, we observe whether there are claims during the year. When claims are present, we have access to the amount of payments for the two types of compulsory coverages: third party liability and personal injury protection (PIP). The liability coverage compensates the policyholder's financial losses arising from property damage or bodily injury of the third party. PIP covers medical expenses for the policyholder in case of a car accident regardless of the legal liability. We consider the combined claims from the two coverage types in this study. See Shi (2016) and Shi and Guszcza (2016) for more detailed description of the Massachusetts automobile insurance data.

In addition to the claim information, the data also contain some basic pricing variables that are commonly used by all insurers, such as driver and vehicle

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

TABLE 1

DESCRIPTIVE STATISTICS OF TOWN-LEVEL OUTCOMES AND COVARIATES.

Variable	Description	Mean	Std	Min	Max
Response Variable					
freq	Frequency of At Least One Claim	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 All-Wheel Drive Vehicle	41.65	9.28	26.93	77.08

characteristics. However, because the dataset represents experiences from several insurance carriers, the amount of policy-level information is very limited and much less than typically used by large carriers that employ advanced analytic techniques. Regarding policyholder characteristics, the only variable available is the age group of the primary driver, i.e., whether the driver is young (age below 25) or senior (age above 65). As for the vehicle characteristics, available variables include vehicle age (in years), vehicle type (passenger car, van, pickup truck and utility vehicle), whether it is a luxury car, and whether it is an all-wheel-drive vehicle. Last, the dataset contains information on the garage town of the insured vehicle. There are in total 351 towns in Massachusetts. The garage town defines the areal units in the statistical modeling of spatial dependency in insurance claims.

In the following analysis, we split the data into two pieces. We use 80% of the data to build the macro (town) level model, and save the rest 20% for validating the microlevel (policy/claim) predictions. Data separation uses stratified sampling based on garage towns. To build a town-level scoring model, we aggregate claims and pricing variables in each garage town. A brief description and the associated summary statistics (mean, standard deviation, minimum and maximum) of aggregated data are exhibited in Table 1. Note that the descriptive statistics reflect the variation of aggregated data across towns. The average exposure is 7,986 car years. The response variable consists of two measurements that correspond to the frequency and severity of claims, respectively. To derive frequency, we count the number of policies with one or more claims over the coverage period and report the corresponding percentage in the table. On average, about 4 out of 100 policies will incur claims over the year, and this probability ranges from 0% to 8% across towns. To define severity, we calculate the average amount of payments per policy conditioning on the occurrence of claims. The

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

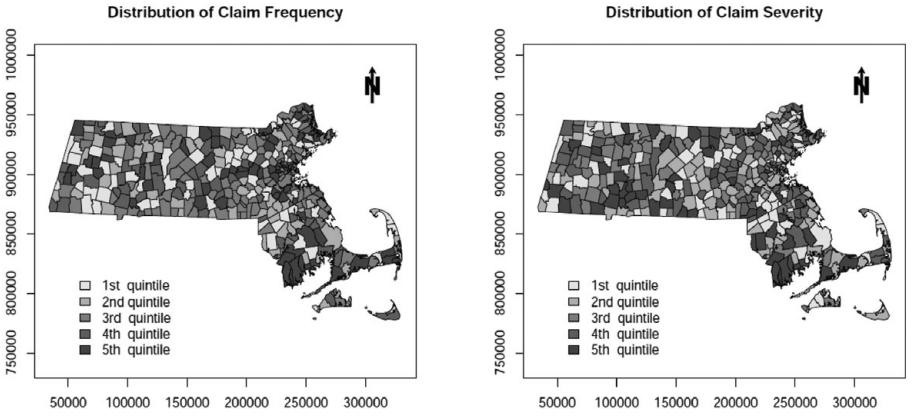


FIGURE 1: Maps of the frequency and severity of automobile insurance claims in Massachusetts.

mean of the (conditional) average payments over towns is about \$3,250. One observes a considerable amount of variation across towns with a minimum of \$680 and a maximum of \$7,300. The covariates are aggregated in a similar manner. We report in the table the percentage for discrete variables and the average for continuous variables. For example, for an average town in Massachusetts, there are about 10% young drivers, about 5% luxury cars and the average vehicle age is 5.4 years. When examining across towns, one also observes significant variability among all explanatory variables.

To motivate the dependent spatial modeling in frequency and severity components, we display in Figure 1 the heat map of the town-level claim probability and average payment amount in Massachusetts. The shape file and the adjacency matrix used for the analysis are obtained from the Office of Geographic Information (MassGIS) of the Government of Massachusetts. The distributions of both variables are demonstrated by their respective quintiles. First, there exists spatial clustering in both frequency and severity components. For instance, see the clusters in the Boston area and in the far southeast of the state. This feature suggests a certain degrees of spatial smoothing in model specification. Second, high (low) frequency areas tend to be associated with high (low) severity areas, indicating the positive dependence between the two components. Third, the spatial variation in the frequency and severity distributions implies heterogeneous degrees of smoothing in the two components.

3. MODELING

3.1. Two-part specification

In the frequency component, we model the number of policyholders in the ith region who have incurred at least one claims, denoted by Y_i^f. We assume that

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

Y_i^f follows a binomial distribution as

$$\begin{aligned} Y_i^f &\sim \text{Bin}(E_i, p_i), \\ f_{Y_i^f}(y) &= \frac{\Gamma(E_i+1)}{\Gamma(y+1)\Gamma(E_i-y+1)} p_i^y (1-p_i)^{E_i-y}, \\ \log\left(\frac{p_i}{1-p_i}\right) &= \boldsymbol{x}_i^f \boldsymbol{\beta}_f + \phi_i^f, \end{aligned} \tag{1}$$

where E_i is the total number of policies written in region i. One can interpret p_i as the probability of one or more claims for a policyholder from region i. A logistic regression is employed for p_i where $\boldsymbol{x}_i^f$ denotes the vector of covariates and $\boldsymbol{\beta}_f$ denotes the associated regression coefficients. The last term ϕ_i^f represents the region-specific effects in the frequency component.

In the severity component, we model the average amount of payments per policy in region i given occurrence of claims, denoted by Y_i^s. A long-tailed regression is employed for Y_i^s due to the skewness and heavy-tails exhibited in the claims distribution. Specifically, we use a log-normal regression as

$$\begin{aligned} Y_i^s &\sim \text{LN}(\mu_i, \tau^{-1}), \\ f_{Y_i^s}(y) &= \frac{\sqrt{\tau}}{y\sqrt{2\pi}} \exp\{-0.5\tau(\ln y - \mu_i)^2\}, \\ \mu_i &= \boldsymbol{x}_i^s \boldsymbol{\beta}_s + \phi_i^s, \end{aligned} \tag{2}$$

with $\text{E}[Y_i^s] = \exp(\mu_i + 0.5\tau^{-1})$ and $\text{Var}[Y_i^s] = \exp(2\mu_i + \tau^{-1})(\exp(\tau^{-1}) - 1)$. As in usual predictive applications, we assume a constant dispersion parameter τ across all geographical regions and perform a regression on the location parameter μ_i. For simplicity, a linear combination of covariates $\boldsymbol{x}_i^s \boldsymbol{\beta}_s$ is employed. Similarly, ϕ_i^s represents counterpart of the region-specific effects in the severity component.

Let $\boldsymbol{\theta}_i^f = (\boldsymbol{\beta}_f, \phi_i^f)$ and $\boldsymbol{\theta}_i^s = (\boldsymbol{\beta}_s, \phi_i^s)$ be the vectors of parameters associated with region i in the frequency and severity components, respectively. Denote function $g(\boldsymbol{\theta}_i^f, \boldsymbol{\theta}_i^s)$ a risk score for region i that could be produced from the model. The score can be used either as a predictor in the models built with more granular level data, or as a rating variable directly for risk classification. Table 2 provides some examples of the risk score.

The first score measures the probability of occurrence of accidents and can be used as a covariate in the frequency modeling in insurance ratemaking. The second score is a variation of the first one and represents the odds ratio of claim occurrence. The third score measures the average size of claims per policy given occurrence and is a more relevant predictor to the severity modeling in ratemaking. The fourth score corresponds to the region-level loss cost that is useful for

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

TABLE 2

EXAMPLES OF THE RISK SCORE.

Example	Score Function
1	$g(\boldsymbol{\theta}_i^f, \boldsymbol{\theta}_i^s) = \mathrm{E}[Y_i^f]/E_i = p_i$
2	$g(\boldsymbol{\theta}_i^f, \boldsymbol{\theta}_i^s) = \mathrm{E}[Y_i^f]/(E_i - \mathrm{E}[Y_i^f]) = p_i/(1-p_i)$
3	$g(\boldsymbol{\theta}_i^f, \boldsymbol{\theta}_i^s) = \mathrm{E}[Y_i^s] = \exp(\mu_i + \frac{1}{2}\tau^{-1})$
4	$g(\boldsymbol{\theta}_i^f, \boldsymbol{\theta}_i^s) = \mathrm{E}[Y_i^f] \times \mathrm{E}[Y_i^s]/E_i = p_i \exp(\mu_i + \frac{1}{2}\tau^{-1})$
5	$g(\boldsymbol{\theta}_i^f, \boldsymbol{\theta}_i^s) = \sqrt{\left(\mathrm{E}[Y_i^{s2}]/\mathrm{E}[Y_i^s]^2\right) / \left(\mathrm{E}[Y_i^f]/E_i\right) - 1} = \sqrt{\exp(\tau^{-1})/p_i - 1}$

the determination of pure premium. The last score is based on the coefficient of variation of the loss cost and is an appropriate risk loading in ratemaking.

In the above examples, it is straightforward to see that the specification for the random effects (ϕ_i^f, ϕ_i^s) is critical for the inference of the score $g(\boldsymbol{\theta}_i^f, \boldsymbol{\theta}_i^s)$. In the next section, we discuss the specification of spatial dependence and cross-sectional dependence in the region-specific random effects.

3.2. Spatial specification

3.2.1. *Review of univariate CAR.* We employ bivariate CAR specifications to accommodate the spatial correlation in the frequency and severity components as well as the correlation between frequency and severity. Below we briefly review the univariate CAR model based on which we introduce the bivariate specification.

The CAR model was originally proposed in Besag (1974). It is the most commonly used approach to account for spatial clustering and variation in areal data, i.e., data collected over geographical units, such as counties, zip codes and so on. Let ψ_i denote the spatial random variable associated with the ith ($i = 1, \ldots, n$) areal unit. Define $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_n)'$. Under the Markov random field assumption, the univariate CAR is defined through conditional distributions:

$$\psi_i | \psi_{-i} \sim N\left(\frac{\lambda}{m_i}\sum_{i\sim j}\psi_j, \frac{1}{\gamma m_i}\right), \; i = 1, \ldots, n, \tag{3}$$

where ψ_{-i} denotes the vector $\boldsymbol{\psi}$ without its ith component, $i \sim j$ indicates that region j is an adjacent neighbor of region i and m_i denotes the number of neighboring regions of the ith areal unit. The model assumes that all neighboring regions are equally influential on every geographical unit. In the above specification, λ is the spatial autocorrelation parameter that controls the spatial dependence among regions, and γ^{-1} is the spatial dispersion parameter.

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

The CAR model (3) corresponds to the following joint distribution:

$$\psi \sim N_n(\mathbf{0}, [\gamma(D - \lambda W)]^{-1}), \tag{4}$$

where N_n denotes the n-dimensional normal distribution, $D = \text{diag}(m_1, \ldots, m_n)$, and W is the adjacency matrix with $W_{ij} = 1$ if geographical units i and j are neighbors and $W_{ij} = 0$ otherwise. A valid joint distribution of ψ requires $1/\zeta_{\min} < \lambda < 1/\zeta_{\max}$, where $\zeta_{\min}$ and $\zeta_{\max}$ are the minimum and maximum eigenvalues of $D^{-1/2}WD^{-1/2}$, respectively. When $\lambda = 0$, model (4) corresponds to the heteroscedastic independence model that induces global smoothing. When $\lambda = 1$, model (4) corresponds to the well-known intrinsic CAR model (Besag *et al.*, 1991) that induces local smoothing by borrowing strength from neighboring regions.

3.2.2. *Bivariate CAR for dependent frequency-severity model.* Let $\phi^f = (\phi_1^f, \ldots, \phi_n^f)'$ and $\phi^s = (\phi_1^s, \ldots, \phi_n^s)'$. To account for the spatial dependence in the random vectors ϕ^f and ϕ^s in the two-part model, a starting point is to assume that each of ϕ^f and ϕ^s follows the univariate CAR framework introduced in Section 3.2.1. For the purposes of our application, it is also important to capture the association between the frequency and severity components. For instance, it is well known that the expectation of the product of frequency and severity is not equal to the product of their expectations if the two components are dependent. Therefore, it is sensible to consider some bivariate version of CAR to join ϕ^f and ϕ^s.

Several bivariate, in general multivariate, CAR models have been proposed in the literature of spatial statistics. The theoretical background on multivariate Gaussian random field is due to Mardia (1988), where the univariate CAR was extended using the separable dependence framework by assuming the identical spatial structure for all variables. The following literature has been focusing on relaxing the restricted separable dependence assumption. Kim *et al.* (2001) proposed a bivariate twofold CAR model to allow for non-separable spatial dependence, though it was not straightforward to extend the framework to higher dimension. Sain and Cressie (2002) presented a formulation of covariance that yields a richer spatial structure, in particular, the asymmetric spatial cross-correlations. However, the approach is computationally expensive. Carlin and Banerjee (2003) and Gelfand and Vounatsou (2003) generalized the CAR model to the multivariate setting using the framework of hierarchical models. For the sake of faster computing, most multivariate CAR priors work with the precision matrix. To improve interpretability, Jin *et al.* (2007) adopted the idea of modeling geostatistical data to the case of lattice data and specified the joint distribution of a multivariate Markov random field directly through conditional and marginal distributions. More recently, Jin *et al.* (2007) proposed an order-free approach to address the issue of arbitrary ordering of variables in the conditional specifications in building multivariate CAR models. Martinez-Beneito

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

(2013) extended the multivariate CAR framework further to incorporate univariate spatial dependencies beyond conditional autoregressive distributions.

In this application, we employ the approach of a linear model of coregionalization. This method has been proposed for multivariate geostatistical data (see, for example, Schmidt and Gelfand, 2003; Gelfand *et al.*, 2004), and recently has also been used for multivariate lattice data (see Jin *et al.*, 2007). Denote $\phi = \left((\phi^f)', (\phi^s)'\right)'$ and define

$$\phi = (B \otimes I_{n\times n})\boldsymbol{v}, \tag{5}$$

where B is a non-singular 2×2 matrix, $I_{n\times n}$ denotes a $n\times n$ identity matrix and $\otimes$ denotes the Kronecker product. Vector $\boldsymbol{v}$ is constructed using $\boldsymbol{v} = \left((\boldsymbol{v}^f)', (\boldsymbol{v}^s)'\right)'$, where $\boldsymbol{v}^f$ and $\boldsymbol{v}^s$, corresponding to the frequency and severity components, respectively, are independent, and each is an $n \times 1$ spatial lattice process. Essentially, one constructs the multivariate spatial process ϕ through linear combinations of univariate processes $\boldsymbol{v}^f$ and $\boldsymbol{v}^s$, and the cross-sectional correlation is introduced by matrix B. When $B = I_{2\times 2}$, model (5) reduces to two independent CAR models. Hence, different assumptions on $\boldsymbol{v}^f$ and $\boldsymbol{v}^s$ will result in different covariance structures in ϕ. We consider two specifications below.

Separable model : Assume that each of $\boldsymbol{v}^f$ and $\boldsymbol{v}^s$ follows a univariate CAR specification (4) and they use the same smoothing parameter:

$$\boldsymbol{v}^f \sim N_n(\mathbf{0}, (D - \lambda W)^{-1}),$$
$$\boldsymbol{v}^s \sim N_n(\mathbf{0}, (D - \lambda W)^{-1}).$$

For identification purposes, we set the spatial dispersion parameters in both $\boldsymbol{v}^f$ and $\boldsymbol{v}^s$ to one. An equivalent strategy is to impose some constraints on matrix B so as to induce unit dispersion. Using (5), it is straightforward to show that

$$\phi \sim N_{2n}(\mathbf{0}, \boldsymbol{\Sigma})$$
$$\boldsymbol{\Sigma} = \begin{pmatrix} (D - \lambda W)\gamma_{11} & (D - \lambda W)\gamma_{12} \\ (D - \lambda W)\gamma_{21} & (D - \lambda W)\gamma_{22} \end{pmatrix}^{-1} \tag{6}$$

Note that $\gamma_{12} = \gamma_{21}$. Let $\Omega^{-1} = \{\gamma_{ij}\}_{i,j=1,2}$, one can also express $\boldsymbol{\Sigma} = \Omega \otimes (D - \lambda W)^{-1}$. The dependence structure is separable in the sense that matrix Ω accommodates the cross-sectional correlation between the frequency and severity components, and $(D - \lambda W)^{-1}$ captures the spatial dependence over geographical regions. Note that matrix B is identifiable up to $BB' = \Omega$.

Non-separable model : One main limitation of the separable specification is that the spatial dependence structure is forced to be identical for both frequency and severity components. This could be a very strong assumption in practice. One simple remedy is to modify the assumptions in $\boldsymbol{v}^f$ and $\boldsymbol{v}^s$. Specifically, assume that both $\boldsymbol{v}^f$ and $\boldsymbol{v}^s$ follow the univariate CAR specification (4) but with

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

different smoothing parameters:

$$\boldsymbol{v}^f \sim N_n(\mathbf{0}, (D - \lambda_f W)^{-1}),$$
$$\boldsymbol{v}^s \sim N_n(\mathbf{0}, (D - \lambda_s W)^{-1}).$$

Under the above assumptions, one can show

$$\phi \sim N_{2n}(\mathbf{0}, \boldsymbol{\Sigma}),$$
$$\boldsymbol{\Sigma} = (B \otimes I_{n\times n}) \begin{pmatrix} (D - \lambda_f W)^{-1} & \mathbf{0} \\ \mathbf{0} & (D - \lambda_s W)^{-1} \end{pmatrix} (B \otimes I_{n\times n})'. \tag{7}$$

When $\lambda_f = \lambda_s = \lambda$, the non-separable model (7) reduces to separable model (6). To see this,

$$\begin{aligned}\boldsymbol{\Sigma} &= \begin{pmatrix} b_{11} I_{n\times n} & b_{12} I_{n\times n} \\ b_{21} I_{n\times n} & b_{22} I_{n\times n} \end{pmatrix} \begin{pmatrix} (D - \lambda W)^{-1} & \mathbf{0} \\ \mathbf{0} & (D - \lambda W)^{-1} \end{pmatrix} \begin{pmatrix} b_{11} I_{n\times n} & b_{12} I_{n\times n} \\ b_{21} I_{n\times n} & b_{22} I_{n\times n} \end{pmatrix}' \\ &= \begin{pmatrix} b_{11}^2 + b_{12}^2 & b_{11}b_{21} + b_{12}b_{22} \\ b_{11}b_{21} + b_{12}b_{22} & b_{21}^2 + b_{22}^2 \end{pmatrix} \otimes (D - \lambda W)^{-1} \\ &= (BB') \otimes (D - \lambda W)^{-1}.\end{aligned}$$

Note that if we define $\Lambda = \text{diag}(\lambda_f, \lambda_s)$, one can also express $\boldsymbol{\Sigma} = (B \otimes I_{n\times n})(I_{2\times 2} \otimes D - \Lambda \otimes W)^{-1}(B \otimes I_{n\times n})'$. Thus, a further extension is to allow Λ to be a symmetric matrix instead of a diagonal matrix. This way one relaxes the independence assumption between $\boldsymbol{v}^f$ and $\boldsymbol{v}^s$ and accommodates the correlation between them via Λ. Under this formulation, the cross-sectional dependence between the frequency and severity will be induced by both B and Λ. For simplicity reason, we will not explore this specification in our study.

3.3. Estimation

We estimate the proposed model using the Markov Chain Monte Carlo (MCMC) method in a Bayesian framework. Let $\boldsymbol{y}^f = (y_1^f, \ldots, y_n^f)'$ and $\boldsymbol{y}^s = (y_1^s, \ldots, y_n^s)'$ denote the data observed from the frequency and severity models, respectively. The posterior distribution of parameters is

$$\begin{aligned}p(\boldsymbol{\beta}_f, \boldsymbol{\beta}_s, \tau, \boldsymbol{v}, \Omega, \Lambda) \propto\ & l(\boldsymbol{y}^f, \boldsymbol{y}^s | \boldsymbol{\beta}_f, \boldsymbol{\beta}_s, \tau, \boldsymbol{v}, \Omega) p(\boldsymbol{v}|\Lambda) p(\Omega) \\ & \times p(\Lambda) p(\boldsymbol{\beta}_f, \boldsymbol{\beta}_s) p(\tau).\end{aligned} \tag{8}$$

The first term corresponds to the likelihood $l(\boldsymbol{y}^f, \boldsymbol{y}^s | \boldsymbol{\beta}_f, \boldsymbol{\beta}_s, \tau, \boldsymbol{v}, \Omega) = l_f(\boldsymbol{y}^f | \boldsymbol{\beta}_f, \boldsymbol{v}, \Omega) l_s(\boldsymbol{y}^s | \boldsymbol{\beta}_s, \tau, \boldsymbol{v}, \Omega)$, where $l_f(\cdot|\cdot)$ and $l_s(\cdot|\cdot)$ are defined by the binomial and lognormal models, respectively. The second term corresponds to the prior of the bivariate spatial process $p(\boldsymbol{v}|\Lambda) = p_f(\boldsymbol{v}^f|\lambda_f) p_s(\boldsymbol{v}^s|\lambda_s)$, where $p_f(\cdot|\cdot)$ and $p_s(\cdot|\cdot)$ are defined by independent CAR models. The rest part corresponds to the prior and hyperprior parameters. The general rule that we follow

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

is to assign vague priors and to make the priors as conjugate as possible. Specifically, smoothing parameters λ_f and λ_s are assumed to follow the uniform distribution on $(1/\zeta_{\min}, 1/\zeta_{\max})$, regression coefficients $\boldsymbol{\beta}_f$ and $\boldsymbol{\beta}_s$ are assumed to follow normal distributions with large variance and dispersion parameter τ is assumed to follow a gamma distribution. The specification of $p(\Omega)$ is discussed below in detail.

In (8), parameter Ω is defined as $\Omega = BB'$. Because of the identification issue in B, we work with Ω directly instead of B. Specifically, we take B as the lower triangular Cholesky decomposition. Recall that Ω captures the correlation between frequency and severity components. To facilitate interpretation, we parameterize Ω in terms of correlation matrix:

$$\Omega = \begin{pmatrix} \sigma_f^2 & \rho\sigma_f\sigma_s \\ \rho\sigma_f\sigma_s & \sigma_s^2 \end{pmatrix}.$$

Therefore, the relation between B and Ω follows:

$$\begin{cases} b_{11} = \sigma_f \\ b_{12} = 0 \\ b_{21} = \rho\sigma_s \\ b_{22} = \sigma_s\sqrt{1-\rho^2}. \end{cases}$$

To set priors, we assume that ρ follows uniform $(-1, -1)$, and $1/\sigma_f$ and $1/\sigma_s$ follow gamma distributions. The above parameterization facilitates prior specifications and model implementation. The approach can be easily generalized to the situation with higher dimension. Note that there is an ordering issue between the frequency and severity components. The particular order that we select above is a natural choice for our application, because the severity component is defined by conditioning on the frequency component in the two-part modeling framework.

4. Simulation

We perform a simple simulation study to investigate the properties of the Bayesian estimates. In the simulation, the frequency and severity outcomes are generated from (1) and (2), respectively, with the following specification:

$$\log(p/(1-p)) = \beta_{f0} + \beta_{f1}X_1 + \beta_{f2}X_2,$$
$$\mu = \beta_{s0} + \beta_{s1}X_1 + \beta_{s2}X_2,$$

where $X_1 \sim$ Uniform(0, 1) and $X_2 \sim$ Bernoulli(0.6). We use only two predictors to keep the model parsimonious, and the same set of explanatory variables are used in the frequency and severity models. Without loss of generality, we set one covariate to be continuous and the other one to be categorical. The spatial variations in (1) and (2) are jointly simulated from (7), where the adjacency matrix

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

TABLE 3

ESTIMATES FROM A SIMULATION WITH 100 REPLICATIONS.

Parameter			Posterior Mean	Relative Bias	Mean Squared Error
β_{f0}	=	−1.50	−1.471	−0.020	0.029
β_{f1}	=	2.00	1.990	−0.005	0.013
β_{f2}	=	−1.00	−1.015	0.015	0.016
β_{s0}	=	7.50	7.517	0.002	0.039
β_{s1}	=	−2.50	−2.490	−0.004	0.053
β_{s2}	=	2.50	2.483	−0.007	0.034
τ	=	15.00	14.360	−0.045	1.433
λ_f	=	0.25	0.248	−0.007	0.039
λ_s	=	0.50	0.451	−0.109	0.104
σ_f	=	15.00	14.430	−0.039	0.740
σ_s	=	10.00	10.921	0.084	1.132
ρ	=	0.75	0.713	−0.053	0.111

from the 351 Massachusetts towns is used in the data generation. The true values of the model parameters as shown in Table 3 are specified to mimic the real insurance application. In the marginal models, exposure E_i in the town-level Massachusetts insurance data described in Section 2 is used in the frequency model, and regression coefficients are specified so that the distributions of p_i and μ_i are close to the real data. The estimated parameters from the Massachusetts data (as reported in Table 5) are used as guidance to determine the dependence parameters in the simulation.

We conduct the simulation in R and we report the results based on 100 replications in Table 3. For each replication, we run the spatially dependent frequency-severity model for 20,000 iterations each with a burn-in of 15,000. The numbers of iteration and burn-in are sufficient to ensure the convergence based on the trace plots from multiple chains. The intercept terms β_{f0} and β_{s0} are assigned flat priors, the slope parameters are assigned normal priors $N(0, 100)$, and the dispersion parameters $(\tau, \sigma_f, \sigma_s)$ are assigned gamma priors Gamma(10, 1). Uniform priors are used for the dependence parameters $(\lambda_f, \lambda_s, \rho)$ on their respective domains, i.e., $\lambda^f, \lambda^s \sim \text{Uniform}(1/\zeta_{\min}, 1/\zeta_{\max})$, and $\rho \sim \text{Uniform}(-1, 1)$. The table shows the average posterior mean, relative bias and mean squared error across the 100 replications. Considering the relatively small sample size (351 areal units), the proposed inference seems sensible to be applied to the Massachusetts automobile insurance data.

5. INFERENCE

To demonstrate the importance of spatial correlation within frequency and severity, as well as the cross-sectional correlation between frequency and severity, we consider alternative model specifications. Specifically, we compare five

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

TABLE 4

GOODNESS-OF-FIT STATISTICS FOR ALTERNATIVE MODELS.

Model	Description	DIC
i	Independent Aspatial Model	9,661
ii	Independent Spatial Model	8,463
iii	Bivariate Intrinsic Spatial Model	8,361
iv	Bivariate Separable Spatial Model	8,323
v	Bivariate Inseparable Spatial Model	8,262

different models: (i) independent aspatial model: This model assumes independence both among town-level responses and between frequency and severity components. Essentially, one separately fits a binomial regression and a lognormal regression without spatial random effects. (ii) Independent spatial model: This model ignores the correlation between the frequency and severity components. The binomial and lognormal models are estimated separably, and in each model, the town-specific effects follow a CAR specification. (iii) Bivariate intrinsic spatial model: This is a nested case of (iv) and (v) where the spatial smoothing parameters in both frequency and severity models are equal to one. But frequency and severity are allowed to be correlated. (iv) Bivariate separable spatial model: This corresponds to specification (6) where the frequency and severity of claims are assumed to follow the same spatial pattern, and the correlation between them is captured by Ω. (v) Bivariate inseparable spatial model: This model uses specification (7) and is the most general model in our study. It allows for different degrees of spatial smoothing in frequency and severity, and in the meanwhile accommodates the cross-sectional correlation.

Bayesian inference for all the above models is implemented in WinBUGS. For each model we run 20,000 MCMC iterations and discard the first 15,000 iterations as burn-in sample. We generate multiple chains from different initial values to assess the convergence of the parameters. Selected diagnostic analysis is reported in the appendix to demonstrate the mixed property of the simulation. For model selection, we turn to the deviance information criterion (DIC) (Spiegelhalter *et al.*, 2002). The DIC is an analogy to the AIC (Akaike information criterion) and BIC (Bayesian information criterion) as a Bayesian model selection criterion. It is calculated as $\text{DIC} = \bar{D} + p_D$, where $\bar{D}$ is the posterior expected deviance and p_D is the effective number of parameters. Both terms are based on the posterior distribution of the deviance statistics which is defined as negative twice the likelihood function plus a constant solely determined by data, i.e., $D(\boldsymbol{\theta}) = -2\log f(\boldsymbol{y}|\boldsymbol{\theta}) + c(\boldsymbol{y})$. It is shown that $\bar{D} = \text{E}_{\boldsymbol{\theta}|\boldsymbol{y}}[D(\boldsymbol{\theta})]$ and $p_D = \bar{D} - D[\text{E}_{\boldsymbol{\theta}|\boldsymbol{y}}(\boldsymbol{\theta})]$. In DIC, a smaller $\bar{D}$ (a larger likelihood) indicates a better fit between the model and the data, and a smaller p_D corresponds to a more parsimonious model, thus a smaller DIC indicates a preferred model. We report the DIC statistics for alternative models in Table 4.

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

TABLE 5

ESTIMATES OF POSTERIOR DISTRIBUTION FOR THE ASPATIAL AND SPATIAL MODELS.

	Independent Aspatial Model		Bivariate Inseparable 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				
λ_f			0.202	(0.009, 0.518)
λ_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)

A comparison between model (i) and model (ii) indicates the significant spatial dependence in the frequency and severity components. Comparisons between models (ii) and (iii)–(v) suggest the importance of cross-sectional correlation between frequency and severity. The DIC statistics select the inseparable bivariate spatial model as the best one. The result emphasizes the advantage of heterogeneous spatial smoothing in the dependent frequency-severity model. Note that DIC is scale-free and is informative only in terms of the ordering of the scores.

For brevity, we present in Table 5 the parameter estimates (posterior mean and 95% credible interval) for the independence aspatial model and the bivariate inseparable spatial model. The reported results correspond to the two extreme cases. One expects the estimation of other specifications falls somewhere in

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

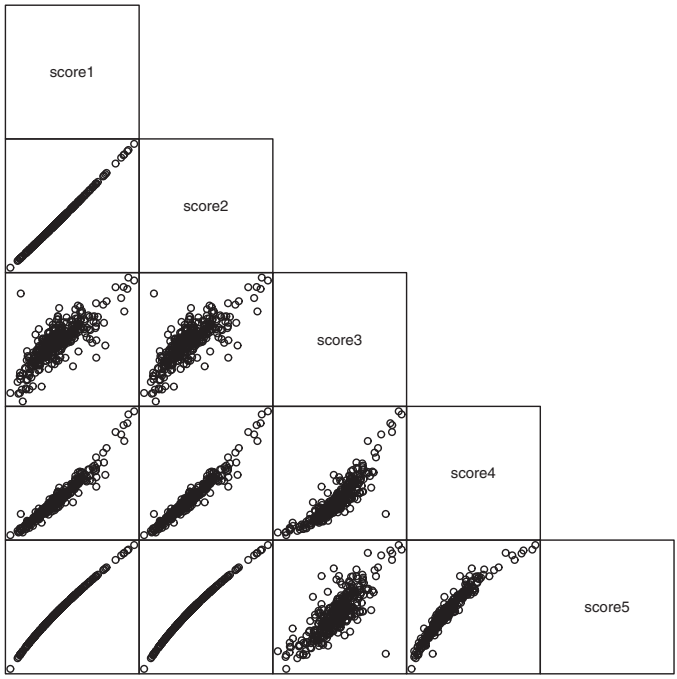


FIGURE 2: Scatter plot matrix of risk scores.

between. Comparing the regression coefficients in frequency and severity components, one finds that there are substantial differences in the posterior means between the two models. However, one should be cautious when interpreting the difference in the posterior distributions. Taking the spatial dependence into consideration does not necessarily lead to more efficient estimates because of the misspecification issue.

The estimates of dependence parameters are quite consistent with exploratory analysis. The spatial dependence in severity (λ_s) appears to be stronger than in frequency (λ_s) for this particular data, providing evidence of heterogeneous spatial variation in the bivariate model. Recall that when $\rho = 0$, we have $B = I_{2\times 2}$, and specification (7) reduces to two independent CAR models. In this sense, parameter ρ reflects the correlation between the frequency and severity components. The strong positive correlation as indicated by the table reinforces the findings in the literature of dependent frequency-severity modeling that is discussed in the introduction. The estimates of both spatial and cross-sectional dependence are in line with observations in Figure 1 and further confirmed by the goodness-of-fit statistics in Table 4.

We conclude this section with a comparison of the risk scores discussed in Section 3.1. We are going to show how to use these scores in the next section. The five risk scores in Table 2 are calculated based on the bivariate inseparable spatial model (the best one in terms of DIC). Figure 2 displays the pair-wise

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

scatter plots of the risk scores that are based on the posterior means. All these scores are positively correlated. In particular, `score2` and `score1` are almost perfectly correlated because of the relatively low claim frequency. The non-linear relation between `score2` and `score5` is due to the specific functional form of our model specification. We denote `score2` as the frequency score (`score.freq`), score3 as the severity score (`score.sev`) and score4 as the pure premium score (`score.pp`). These three scores will be used in the following applications.

Note that using MCMC iterations, one could derive the posterior distribution of each score in each town. To emphasize this, Figure 3 shows the posterior distributions of the frequency, severity and pure premium scores for all Massachusetts towns. The towns are ordered based on the posterior means of the pure premium score. As shown in Figure 2, we observe the positive relation between the frequency and severity scores.

6. Applications

This section discusses two applications of the model output (town-level risk scores) in insurance operations. The first one demonstrates that the town-level risk scores can be used to improve micro-level prediction. For simplicity, we focus on the posterior means of the risk scores. The second one shows that the analysis of aggregated data provides a natural setting for risk clustering. In this application, we will demonstrate the usage of score distributions in market segmentation.

6.1. Ratemaking

Ratemaking is a classical actuarial problem where one wants to determine the price of an insurance contract. We are examining a simpler version of the problem, pure premium determination. The quantity of interest is the expected cost of an individual contract. Two classes of methods are commonly used for pricing non-life insurance in practice, the two-part model and the pure premium model (Frees *et al.*, 2014). We show that that the town-level risk scores help improve both in-sample goodness-of-fit and out-of-sample prediction. Recall that 20% of data are reserved for application. Imagine that the data used to develop risk scores are from the insurance market. The insurer wants to incorporate the risk score into its own claim experience. We further separate this data into two pieces, using 75% to build pricing models and 25% to test the predictive performance.

Table 6 shows the summary statistics of the individual level data. It is noticeable that policyholders with more than one claims account for less than 1% of the portfolio. This is partly because we only considered the minimum required coverage for drivers in Massachusetts. As anticipated, the claim severity are right skewed and heavy-tailed.

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

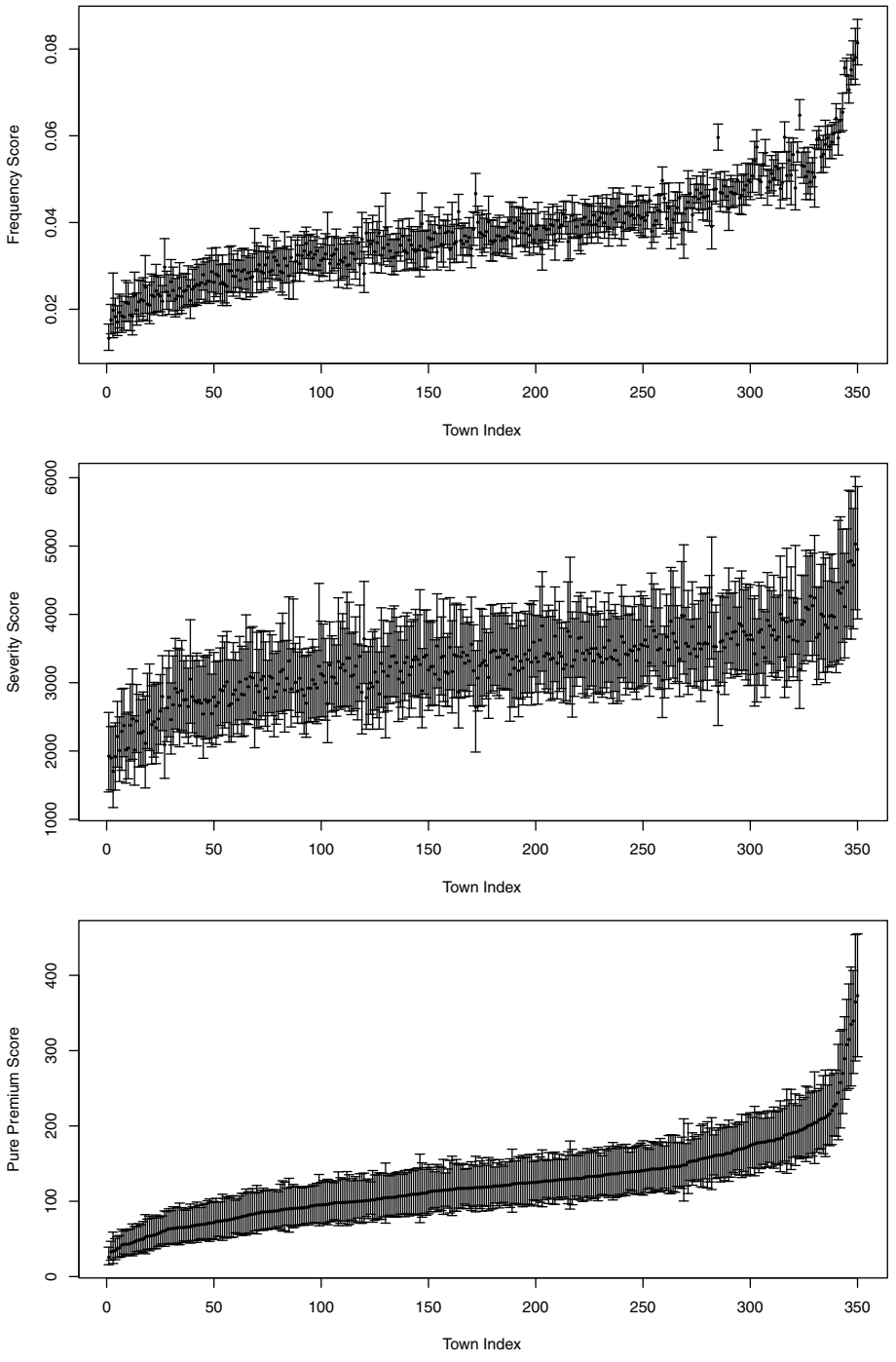


FIGURE 3: Credible intervals for frequency score, severity score and pure premium scores.

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

TABLE 6

DESCRIPTIVE STATISTICS OF POLICY-LEVEL OUTCOMES AND COVARIATES.

freq	0	1	2	3	4
Percentage	95.45	4.34	0.20	0.01	0.00
	Mean	Std			
size	3804.05	6673.86			
young	9.86				
senior	14.19				
vehage	5.42	2.98			
lux	5.04				
van	7.98				
pickup	10.34				
utility	23.83				
awd	35.49				

We consider two versions of the frequency-severity model. The first one examines claims at policy level. A logistic regression is used to model the probability of a policy having one or more claims, and a lognormal regression is used to model the total amount of payments from a policy given occurrence of claims. The second one examines claims at claim level. A Poisson regression is employed to model the number of claims of a policyholder, and a gamma regression (with log link) is employed to model the amount of individual claims assuming independence among multiple claims from the same policyholder. Though not reported here, we also explored more flexible models (negative binomial regression and zero-inflated Poisson regression) for claim counts, and did not find the results superior to the Poisson model. It is not surprising given the fact that over 99% policyholders have either zero or one claim over the year.

To examine the predictive lift due to the scores from the spatial model, we include `score.freq` in the frequency models, and `score.sev` in the severity models. Note that all other predictors are measured at policyholder level. Estimation results are summarized in Table 7. Since the statistical significance of policy-level covariates is not of our primary interest, we skip the discussion of their effects in this study. The effects of both frequency and severity scores are positive and statistically significant, suggesting that a high frequency score is associated with a larger probability of claims, and a high severity score is associated with a larger amount of payments, other things equal. We will also see this significant effect in the out-of-sample prediction.

The pure premium model is based on the Tweedie generalized linear model. We perform a regression on the location parameter only. We consider three different ways to include the town-level risk score. First, if one thinks of the pure premium score (`score.pp`) as an estimate of the cost of a contract without taking into account individual risk factors, it is sensible to use the score as an offset variable in the Tweedie model. Second, to extend this idea, one could also

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

TABLE 7

ESTIMATION RESULTS OF FREQUENCY-SEVERITY MODELS.

Two-Part: Policy Data					
Logit	Estimate	Std. Error	LN	Estimate	Std. Error
Intercept	0.148	0.088	Intercept	4.746	0.805
young	0.632	0.021	young	0.198	0.034
senior	0.032	0.023	senior	0.023	0.038
vehage	0.005	0.003	vehage	−0.004	0.004
lux	−0.287	0.041	lux	−0.040	0.068
van	0.082	0.029	van	0.015	0.048
pickup	0.209	0.030	pickup	0.279	0.052
utility	0.180	0.028	utility	0.161	0.049
awd	−0.009	0.026	awd	0.052	0.045
score.freq	1.102	0.029	score·	0.288	0.099
Two-Part: Claim Data					
Poisson	Estimate	Std. Error	Gamma	Estimate	Std. Error
Intercept	0.150	0.084	Intercept	3.425	0.831
young	0.613	0.020	young	0.146	0.035
senior	0.029	0.022	senior	−0.079	0.039
vehage	0.005	0.002	vehage	−0.005	0.004
lux	−0.288	0.039	lux	0.041	0.070
van	0.079	0.027	van	−0.075	0.049
pickup	0.217	0.029	pickup	0.246	0.054
utility	0.178	0.027	utility	0.152	0.051
awd	−0.009	0.025	awd	−0.080	0.046
score.freq	1.103	0.027	score·sev	0.581	0.102

estimate the effect of the score from the data. In this sense, we simply use `score.pp` as a predictor in the model. Third, instead of combining two scores `score.freq` and `score.sev` into the pure premium score, we use them as covariates and examine their effects separately. The estimation results are summarized in Table 8. The estimate of regression coefficient is 0.957, indicating that the pure premium score is a good estimate of the insurance cost. As anticipated as well, both frequency score and severity score are positively associated with loss cost.

The predictive performance is evaluated using out-of-sample validation. Considering the non-continuous nature of the outcome variables, we employ Gini index as the validation statistic. See Frees *et al.* (2011) for the theoretical background of Gini index and Frees *et al.* (2014) for its applications in insurance. The Gini correlation and the simple Gini of alternative specifications are presented in Table 9. The validation statistics are calculated for frequency, severity and pure premium models separately. Within each group, we consider three different sets of predictors: (1) Policy-level characteristics only: this set ignores

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

TABLE 8

ESTIMATION RESULTS OF PURE PREMIUM MODELS.

	Pure Premium				Frequency-Severity	
Score	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Intercept	−0.086	0.038	0.134	0.225	0.789	1.770
young	0.761	0.045	0.759	0.045	0.759	0.045
senior	−0.057	0.047	−0.059	0.047	−0.059	0.047
vehage	−0.001	0.005	−0.001	0.005	−0.001	0.005
lux	−0.258	0.078	−0.258	0.078	−0.258	0.078
van	−0.027	0.060	−0.027	0.060	−0.026	0.060
pickup	0.412	0.061	0.409	0.061	0.412	0.061
utility	0.297	0.058	0.300	0.058	0.301	0.058
awd	−0.072	0.053	−0.075	0.053	−0.076	0.053
score.pp	1.000	–	0.957	0.044	–	–
score.freq				0.950	0.091	
score.sev				0.868	0.190	

the town-specific effects. (2) Policy information and town-specific heterogeneity. Thinking of Massachusetts as population, we add a town-level fixed effect. (3) Policy information and the corresponding risk score. One can think of risk scores as spatially smoothed town effects. It is not surprising to see the predictive lift added by the geographical location of garage towns of the policyholder. We further observe the additional lift due to risk scores when comparing with town fixed effects. Another advantage of using risk scores is the significant reduction in the computational time. In the pure premium panel, the validation statistics are almost the same regardless of whether using `score.pp` as an offset or a regressor. This is because the estimated regression coefficient of `score.pp` in the Tweedie model is very close to one. In the last two rows of the table, we report the Gini statistics for the two versions of the two-part model so as to compare their performance with the pure premium model. The performance is comparable, especially the Poisson-gamma model. This result is sensible because the Tweedie random variable is a Poisson sum of gamma variables.

6.2. Market segmentation

The town-level risk scores produced by the bivariate spatial model can be used in risk clustering for market segmentation purposes. As above, if one thinks of the scores as being produced using the market data (pooled from the experiences of several insurers), the insurer could use the score to segment risks geographically, which is useful for marketing and underwriting. In marketing, the insurer wants to locate the potential profitable customers. In underwriting, the insurer decide the riskiness of the policyholder. For new customers, the risk score could serve as a priori information about the policyholder's risk level. For existing customers,

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

TABLE 9

OUT-OF-SAMPLE VALIDATION FOR THE FREQUENCY-SEVERITY AND PURE PREMIUM MODELS.

	Gini Correlation	Simple Gini
Frequency Model		
Logit: policy info only	3.501	9.160
Logit: policy info + town	6.000	15.702
Logit: policy info + score· freq	6.517	17.053
Poisson: policy info only	3.526	9.452
Poisson: policy info + town	6.108	16.375
Poisson: policy info + score.freq	6.571	17.615
Severity Model		
LN: policy info only	2.668	2.715
LN: policy info + town	4.703	4.787
LN: policy info + score.sev	5.235	5.328
Gamma: policy info only	4.190	4.261
Gamma: policy info + town	5.554	5.649
Gamma: policy info + score.sev	6.128	6.233
Pure Premium		
Tweedie GLM: policy info only	2.467	13.319
Tweedie GLM: policy info + town	4.015	21.681
Tweedie GLM: policy info + score.pp (offset)	4.287	23.152
Tweedie GLM: policy info + score.pp	4.287	23.153
Tweedie GLM: policy info + score.freq + score.sev	4.289	23.160
Two-part: Logit (`score.freq`) + LN (`score.sev`)	4.184	22.597
Two-part: Poisson (`score.freq`) + Gamma (`score.sev`)	4.288	23.155

the risk score is a supplement to the insurer's own claim experience, which is valuable especially for small insurers without credible data.

In this study, we demonstrate the town-level risk clustering based on the entire distribution of the risk score. We employ the k-medoids algorithm for clustering that is implemented using partitioning around medoids (PAM) (Theodoridis and Koutroumbas, 2003). To be self-contained, we provide a detailed description of the clustering technique in the appendix. The essential idea is to partition points into groups such that within each group the distance between observations and representative objects (known as medoids) is minimized. The advantage of the PAM is that it supports any arbitrary distance measure and it is more robust than k-means algorithm. This feature is particularly important to our application. Recall that from the output of MCMC iterations one obtains the entire distribution of the risk score at each town rather than a single point estimate. In the clustering analysis, we use the Jensen–Shannon divergence (JSD) to measure the distance between two score distributions (Manning and Schütze, 1999). The optimal number of clusters is determined using the Silhouette coefficient (Rousseeuw, 1987).

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

We perform the clustering analysis for the frequency, severity and pure premium scores. Two comparisons are reported. First, we compare the clustering results that is based on the entire distribution of the score and that is based on a limited amount of information of the score. In the former, the similarity between two scores are measured using the JSD. In the latter, we cluster using only two features of each score, the mean and the variance (both are normalized), and the similarity is based on their Euclidean distance. For visualization, we plot clusters on the mean-variance dimension. The results are summarized in Figure 4. In general, the distribution-based approach is expected to group data into more clusters than the mean-variance based approach. This is indeed what is observed for the frequency score. The clustering results for the severity score from the two methods are very close. This is explained by the small variability in the variance of the score as exhibited in the figure. The pure premium score is a combination of frequency and severity, and the clustering result reflects this fact. Second, we compare the clustering using the scores from the spatial model and the scores from the aspatial model. Both are based on the score distribution, and the optimal clustering is visualized in Figure 5. As anticipated, there are fewer clusters when we allow some degrees of local smoothing among the risk scores over geographic regions, and the associated map of risk segmentation is much more smooth when using scores from the spatial model compared with the aspatial model.

7. Conclusion

In this paper, we proposed a spatially dependent frequency-severity modeling framework for risk classification in non-life insurance. Specifically, we extended the standard two-part model by (1) allowing the frequency and severity components to be correlated, and (2) incorporating the correlation over geographical regions in each component. To achieve this purpose, we adopted the idea of linear models of co-regionalization. In the application, we demonstrated the importance of both types of dependencies in insurance claims modeling.

Our study fits in the broader literature of spatial statistics. We employed the CAR-based model, which is a commonly used strategy for areal data, to account for spatial dependence in the town-level aggregated insurance claims. An alternative strategy is to consider the methods for handling geostatistical data. In each case, one wants to balance the computation and interpretability when modeling the covariance matrix.

From the perspective of application, a relevant question of practical interest is how to use macro-level information to improve micro-level prediction. In this application, we aggregated data at town- (macro) level and calibrated the spatial statistical model using the aggregated data. We further showed that the model output could help improve the prediction at policy- and claim- (micro) level observations. As an alternative, one could build a mixed effects model with town-level heterogeneity being a spatial process. The downside is the computational complexity especially when the sample size is large.

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

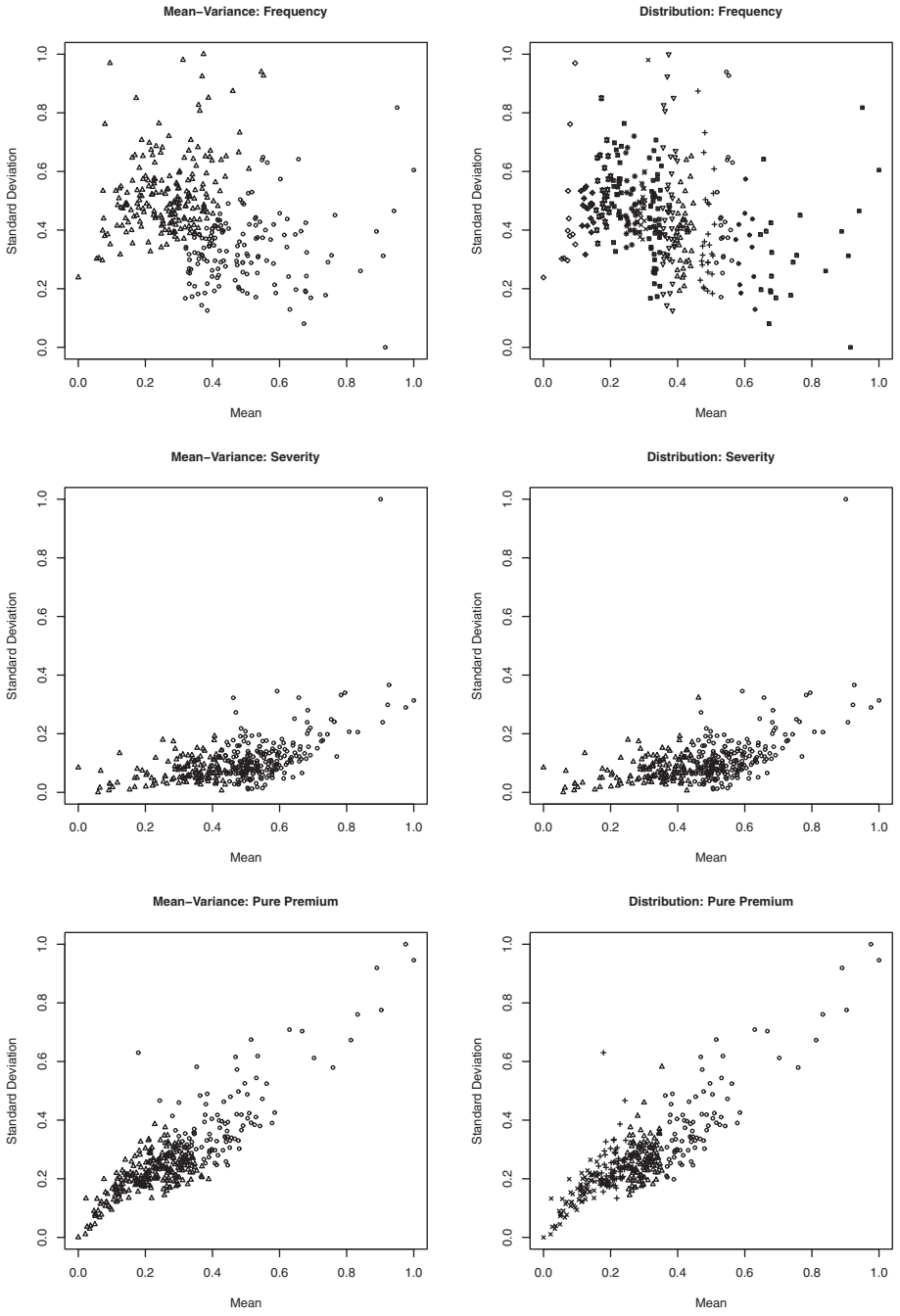


FIGURE 4: Clusters in the frequency, severity and pure premium scores on the mean-variance space. The left panel groups data based on the mean and variance of the score and the right panel groups data based on the score distribution.

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press



FIGURE 5: Clusters in the frequency, severity and pure premium scores on map. The left panel uses scores from the aspatial model and the right panel uses scores from the spatial model.

ACKNOWLEDGEMENTS

Peng Shi thanks the support by the Centers of Actuarial Excellence (CAE) Research Grant from the Society of Actuaries and the Wisconsin Naming Partners Fellowship from the University of Wisconsin-Madison.

REFERENCES

BESAG, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, **36**(2), 192–236.

BESAG, J., YORK, J. and MOLLIÉ, A. (1991) Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43**(1), 1–20.

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

BRECHMANN, E. and CZADO, C. (2014) Spatial modeling. In *Predictive Modeling Applications in Actuarial Science: Volume II, Predictive Modeling Techniques* (eds. E.W. Frees, G. Meyers and R.A. Derrig), pp. 260–279. Cambridge: Cambridge University Press.

CARLIN, B.P. and BANERJEE, S. (2003) Hierarchical multivariate CAR models for spatio-temporally correlated survival data. In *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting* (eds. J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith and M. West), pp. 45–63. New York, NY: Oxford University Press.

CZADO, C., KASTENMEIER, R., BRECHMANN, E.C. and MIN, A. (2012) A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal*, **2012**(4), 278–305.

CZADO, C., SCHABENBERGER, H. and ERHARDT, V. (2014) Non nested model selection for spatial count regression models with application to health insurance. *Statistical Papers*, **55**(2), 455–476.

DIMAKOS, X.K. and DI RATTALMA, A.F. (2002) Bayesian premium rating with latent structure. *Scandinavian Actuarial Journal*, **2002**(3), 162–184.

ERHARDT, V. and CZADO, C. (2012) Modeling dependent yearly claim totals including zero claims in private health insurance. *Scandinavian Actuarial Journal*, **2012**(2), 106–129.

FREES, E.W., MEYERS, G. and CUMMINGS, A.D. (2011) Summarizing insurance scores using a Gini index. *Journal of the American Statistical Association*, **106**(495), 1085–1098.

FREES, E.W., MEYERS, G. and CUMMINGS, A.D. (2014) Insurance ratemaking and a Gini index. *Journal of Risk and Insurance*, **81**(2), 335–366.

GELFAND, A.E., SCHMIDT, A.M., BANERJEE, S. and SIRMANS, C. (2004) Nonstationary multivariate process modeling through spatially varying coregionalization. *Test*, **13**(2), 263–312.

GELFAND, A.E. and VOUNATSOU, P. (2003) Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, **4**(1), 11–15.

GSCHLÖßL, S. and CZADO, C. (2007) Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal*, **2007**(3), 202–225.

JIN, X., BANERJEE, S. and CARLIN, B.P. (2007) Order-free co-regionalized areal data models with application to multiple-disease mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**(5), 817–838.

KIM, H., SUN, D. and TSUTAKAWA, R.K. (2001) A bivariate Bayes method for improving the estimates of mortality rates with a twofold conditional autoregressive model. *Journal of the American Statistical Association*, **96**(456), 1506–1521.

KRÄMER, N., BRECHMANN, E.C., SILVESTRINI, D. and CZADO, C. (2013) Total loss estimation using copula-based regression models. *Insurance: Mathematics and Economics*, **53**(3), 829–839.

MANNING, C.D. and SCHÜTZE, H. (1999) *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.

MARDIA, K. (1988) Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. *Journal of Multivariate Analysis*, **24**(2), 265–284.

MARTINEZ-BENEITO, M.A. (2013) A general modelling framework for multivariate disease mapping. *Biometrika*, **100**(3), 539–553.

ROUSSEEUW, P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics,* **20**, 53–65.

SAIN, S.R. and CRESSIE, N. (2002) Multivariate lattice models for spatial environmental data. *Proceedings of Joint Statistical Meetings: Sections on Statistics and the Environment*, 2820–2825.

SCHMIDT, A.M. and GELFAND, A.E. (2003) A Bayesian coregionalization approach for multivariate pollutant data: Application of recent advances in space-time statistics to atmospheric data. *Journal of Geophysical Research,* **108**(D24), STS10.1–STS10.8.

SHI, P. (2016) Insurance ratemaking using a copula-based multivariate Tweedie model. *Scandinavian Actuarial Journal*, **2016**(3), 198–215.

SHI, P., FENG, X. and IVANTSOVA, A. (2015) Dependent frequency–severity modeling of insurance claims. *Insurance: Mathematics and Economics*, **64**, 417–428.

SHI, P. and GUSZCZA, J. (2016) Frameworks for general insurance ratemaking: Beyond the generalized linear model. In *Predictive Modeling Applications in Actuarial Science: Volume II, Predictive Modeling Techniques* (eds. E.W. Frees, G. Meyers and R.A. Derrig), pp. 100–125. Cambridge: Cambridge University Press.

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

SPIEGELHALTER, D.J., BEST, N.G., CARLIN, B.P. and VAN DER LINDE, A. (2002) Bayesian measures of model complexity and fit (with discussion) *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(4), 583–639.
THEODORIDIS, S. and KOUTROUMBAS, K. (2003) *Pattern Recognition*. San Diego, CA: Academic Press.

PENG SHI
Wisconsin School of Business
University of Wisconsin–Madison
Madison, WI 53706, USA
E-Mail: pshi@bus.wisc.edu

KUN SHI
Deloitte Consulting
2200 Ross Ave #1600, Dallas, TX 75201
E-Mail: kushi@deloitte.com

APPENDIX A

A.1. CLUSTERING ANALYSIS

This section briefly describes three key components of the clustering technique employed in the application. The first is the PAM algorithm that is used to implement the k-medoids method. There are three steps in the algorithm: (1) initialization: randomly select a subset of n towns as medoids; (2) assignment: assign each town to its closest medoid; (3) update: calculate the average distance between an observation and all other observations in the same cluster, and set the observation with smallest average distance as the new medoid. Repeat these three steps until convergence.

The second is the distance measure between two distributions, the JSD. The statistics between two distributions g_1 and g_2, denoted by $\mathrm{JSD}(g_1, g_2)$, is defined as

$$\mathrm{JSD}(g_1, g_2) = \frac{1}{2}\mathrm{KLD}(g_1, \bar{g}) + \frac{1}{2}\mathrm{KLD}(\bar{g}, g_2)$$

where $\bar{g} = (g_1 + g_2)/2$ and $\mathrm{KLD}(\cdot, \cdot)$ denote the Kullback–Leibler divergence between two distributions.

The third is the Silhouette coefficient used to determine the optimal number of clusters. The Silhouette coefficient for the score in the ith town, denoted by $S(i)$, is defined as

$$S(i) = \frac{d(i) - c(i)}{\max\{c(i), d(i)\}}.$$

Here, $c(i)$ is the average distance between town i and all other towns in the same cluster, and $d(i)$ is average distance of town i to all towns in the closest other cluster. By definition, $S(i)$ is bounded by -1 and 1. A smaller $c(i)$ indicates better cluster cohesion and a larger $d(i)$

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

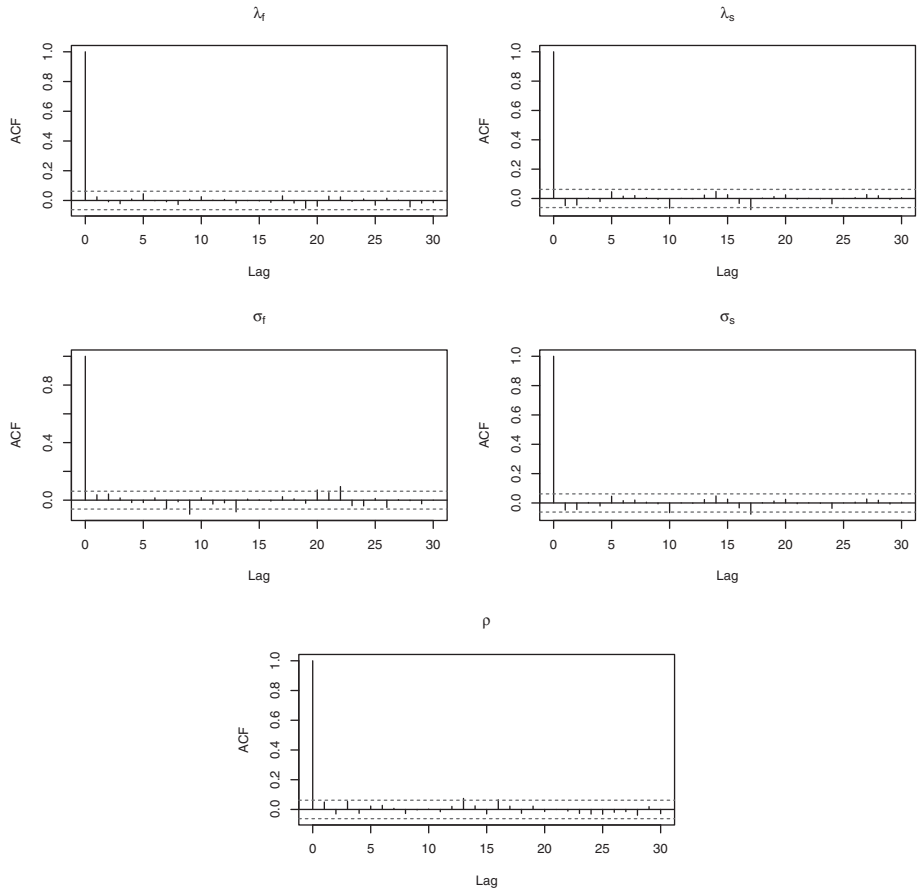


FIGURE A1: Autocorrelation of MCMC iterations for the dependence parameters.

indicates better cluster separation. An observation which is much closer to its own cluster than to any other cluster has a high positive $S(i)$ value. An $S(i)$ close to zero implies that the observation lies somewhere between two clusters while a large negative $S(i)$ value indicates that the observation is assigned to the wrong cluster. The overall Silhouette coefficient is the average over all towns $\bar{S} = \sum_{i=1}^{n} S(i)/n$.

A.2. DIAGNOSTICS OF POSTERIOR DISTRIBUTIONS

We perform routine diagnostic analysis to examine the mixing properties of the Markov chains. For brevity, we only report some important results in this section. The first concern is regarding the temporal dependence in the parameters. Figure A1 displays the autocorrelation of the MCMC iterations for the dependence parameters in the spatial model. The marginal parameters show similar pattern and the results suggest no strong temporal dependence in the model parameters. The second analysis examines the effects of initial values on the convergence and hence ensure the burn-in sample to be large enough. Because of model complexity,

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

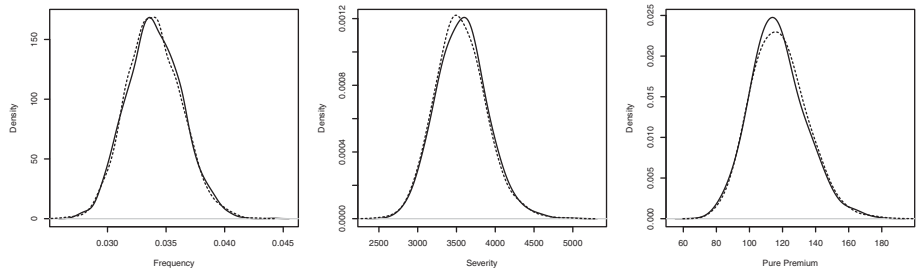


FIGURE A2: Posterior distributions of the frequency, severity and pure premium scores for an average town from multiple Markov chains.

instead of focusing on the posterior distributions of parameters, we present in Figure A2 the posterior distributions of the frequency, severity and pure premium scores from two chains. The consistency between multiple chains indicates the robustness of the estimates.

A.3. SUPPLEMENTARY MATERIALS

This section provides more details on the Tweedie generalized linear model and Gini index. The Tweedie distribution is defined as a Poisson sum of gamma random variables. Specifically, let Y_i denote the per exposure amount of claims for risk i. We can express Y_i as $Y_i = (Z_{i1} + \cdots + Z_{iN_i})/\omega_i$, where N_i is the number of claims, Z_{ij} $(j = 1, \ldots, N_i)$, independent of N_i, are an i.i.d. sequence indicating the corresponding amount of claims and ω_i denotes the exposure. The Tweedie compound Poisson model is derived by assuming that N_i is Poisson distributed with mean $\omega_i\lambda_i$ and Z_{ij} is gamma distributed with shape parameter α and scale parameter γ_i.

The Tweedie variable Y_i has a mass probability at zero:

$$f_i(0) = \Pr(Y_i = 0) = \Pr(N_i = 0) = e^{-\omega_i\lambda_i}.$$

Notice that Y_i conditional on $N_i = n$ follows a gamma distribution with shape $n\alpha$ and scale γ_i/ω_i, one could show the density of the Tweedie distribution for $y > 0$ as

$$f_i(y) = \sum_{n=1}^{\infty} e^{-\omega_i\lambda_i} \frac{(\omega_i\lambda_i)^n}{n!} \frac{1}{(\gamma_i/\omega_i)^{n\alpha}} y^{n\alpha-1} e^{-y\gamma_i/\omega_i}.$$

The Tweedie distribution is a special case of exponential dispersion models, a class of models used to describe the random component in generalized linear models. To see this, consider the following reparameterizations:

$$\lambda_i = \frac{\mu_i^{2-p}}{\phi(2-p)}, \quad \alpha = \frac{2-p}{p-1}, \quad \gamma_i = \phi(p-1)\mu_i^{p-1}. \tag{9}$$

With the above relationships, one can show that the distribution of Y_i is

$$f_i(y) = \exp\left[\frac{\omega_i}{\phi}\left(\frac{-y}{(p-1)\mu_i^{p-1}} - \frac{\mu_i^{2-p}}{2-p}\right) + S(y;\phi/\omega_i)\right],$$

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press

where

$$S(y;\phi/\omega_i)=\begin{cases}0 & \text{if } y=0\\ \ln\sum_{n\geq 1}\left\{\frac{(\omega_i/\phi)^{1/(p-1)}y^{(2-p)/(p-1)}}{(2-p)(p-1)^{(2-p)/(p-1)}}\right\}^n\frac{1}{n!\Gamma(n(2-p)/(p-1))y} & \text{if } y>0.\end{cases}$$

Hence, the distribution of Y_i belongs to the exponential family with parameters μ_i, ϕ, and $p\in(1,2)$, and we have

$$\mathrm{E}(Y_i)=\mu_i \quad \text{and} \quad \mathrm{Var}(Y_i)=\frac{\phi}{\omega_i}\mu_i^p.$$

It is also worth mentioning the two limiting cases of the Tweedie model: $p\to 1$ results in the over-dispersed Poisson model and $p\to 2$ results in the gamma GLM. The Tweedie compound Poisson model accommodates the situations inbetween. To perform regression, one can connect mean with covariates through a link function $g_\mu(\mu_i)=\boldsymbol{x}_i'\boldsymbol{\beta}$, where $\boldsymbol{x}_i$ is the vector of covariates and $\boldsymbol{\beta}$ is the vector of corresponding regression coefficients.

The ordered Lorenz curve and the associated Gini index are suitable for out-of-sample validation where the outcome distribution is featured with high proportion of zeros and the skewness and heavy tails. Using insurance as an example, the essential idea of the ordered Lorenz curve is to measure the discrepancy between the premium and loss distributions. Let $B(\boldsymbol{x})$ be the base premium and $P(\boldsymbol{x})$ be the competing premium, both depending on a set of exogenous variables $\boldsymbol{x}$. The ordered premium and loss distributions are defined based on the relativity $R(\boldsymbol{x})=P(\boldsymbol{x})/B(\boldsymbol{x})$ as

$$\hat{H}_P(s)=\frac{\sum_{i=1}^n B(\boldsymbol{x}_i)I(R(\boldsymbol{x}_i)\leq s)}{\sum_{i=1}^n B(\boldsymbol{x}_i)} \quad \text{and} \quad \hat{L}_P(s)=\frac{\sum_{i=1}^n y_i I(R(\boldsymbol{x}_i)\leq s)}{\sum_{i=1}^n y_i}.$$

The ordered Lorenz curve is the plot of $\left(\hat{H}_P(s),\hat{L}_P(s)\right)$. The 45-degree line, known as the line of equality, indicates the percentage of losses equals the percentage of premiums. The associated Gini index is defined as twice the area between the ordered Lorenz curve and the line of equality, and it may range over $(-1,1)$. A curve below the line of equality suggests that the insurer could look to the competing premium to identify more profitable contracts. In this study, we employ two related statistics, the Gini coefficient and the simple Gini. The former measures the correlation between the predicted outcome and the rank of predictor, and the latter is twice the average covariance between the predicted outcome and the rank of the predictor. For both cases, a larger statistic indicates a better prediction.

https://doi.org/10.1017/asb.2017.7 Published online by Cambridge University Press