Geographical Ratings with Spatial Random Effects in a Two-Part Model

by Chun Wang, Elizabeth D. Schifano and Jun Yan

ABSTRACT

Rating areas are commonly used to capture unexplained geographical variability of claims in insurance pricing. A new method for defining rating areas is proposed using a two-part generalized geoadditive model that models spatial effects smoothly using Gaussian Markov random fields. The first part handles zero/nonzero expenses in a logistic model; the second handles nonzero expenses (on log-scale) in a linear model. Both models are fit with R package INLA for Bayesian inferences. The resulting spatial effects are used to construct more representative ratings. The methodology is illustrated with simulated data based on zipcode areas, but modeled on zipcode- or county-level.

KEYWORDS

Geoadditive model; INLA; rating area; territory analysis

1. Introduction

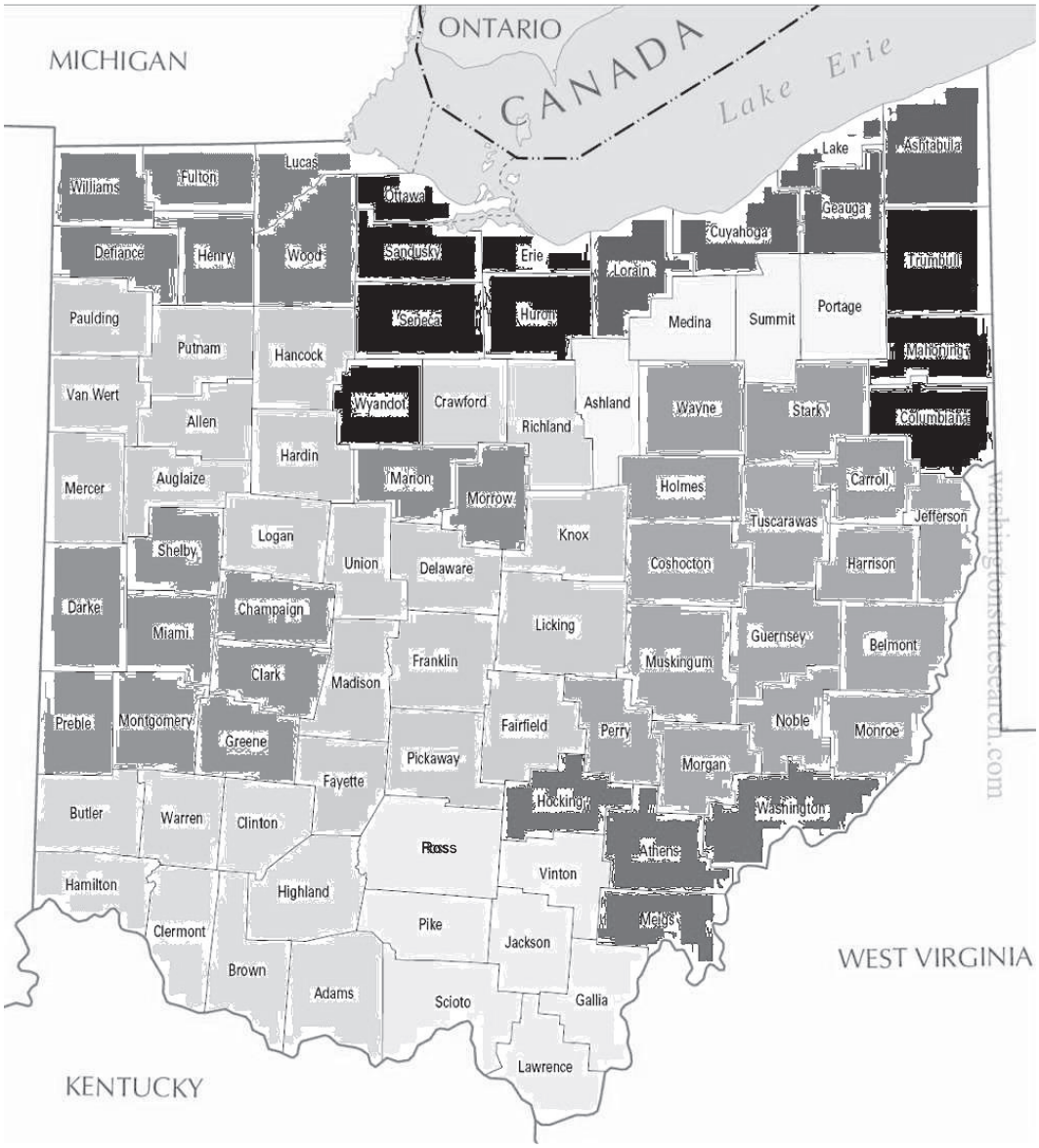
Geographical variability of claims in health insurance is a well-known issue that is important in pricing insurance products. The relevant research is known as geographical ratemaking in property and casualty insurance (e.g., Griz 2015; McClenahan 1990). Area of residence is one of the factors that health insurers can use when adjusting premiums by the Patient Protection and Affordable Care Act, as it helps to account for the spatial variability that cannot be explained by other factors such as age and smoking status (e.g., Kofman and Pollitz 2006; NAIC and CIPR 2011). A common way to handle the spatial variation is by clustering small geographic regions (e.g., at the county level or zipcode level) into larger areas, each of which has its own rate in product pricing (e.g., Werner and Modlin 2010). These grouped areas are usually called geographic rating areas. Although geographic rating is permitted in most states, it is limited by state/federal regulations to the use of the first three digits of zipcode or county (https://www.cms.gov/cciio/programs-and-initiatives/health-insurance-market-reforms/state-gra.html). For example, Figure 1 is a sample rating area map of Ohio approved by the Ohio Department of Insurance, where 88 counties in the state are grouped into 17 rating areas.

Rating areas need to be defined with actuarial justifications. Early rating areas were defined based on subjective information such as agent feedback or loss ratios that may have lacked credibility; some historical territory definitions lacked statistical support and may have lost meaning over time (Jennings 2008, p.34). Modern clustering methods, non-model based or model based, have been applied to residuals from standard models such as generalized linear models (GLMs), (Haberman and Renshaw 1996; McCullagh 1984) to group basic geographic units such as counties into clusters based on historical experience, modeled experience, or well-defined similarity rules (Werner and Modlin 2010; Yao 2008). A unique requirement of the cluster analysis in this context is that there are various social and regulatory acceptability constraints such as the minimum area or geographic contiguity of each territory (Weibel and Walsh 2008). The homogeneity of risk classification in a rating area can be assessed by the within cluster variance as a percentage of the total variance (Jennings 2008; Miller 2004). Rating areas with territorial boundaries from clustering methods are coarse, as shown in Figure 1, in that: (1) units within one rating area may not be that homogeneous; and (2) neighboring units across a boundary may not be that different. Further, clustering methods cannot handle geographic units with no observations, which is not uncommon in less populated areas.

Because of the spatial contiguity and the "spill-over" effect, it is reasonable to assume that the geographic risk surface is smooth. With the aid of geographic information systems, local smoothing and inverse distance weighted averages have been used to define ratings at the basic geographical unit level without territorial boundaries (Brubaker 1996; Christopherson and Werland 1996). These methods are often applied to the residuals from models that account for large scale variation explained by other predictors. Therefore, the smoothing and the modeling are in two separate stages, which makes inference on the ratings difficult and often disregarded. A modeling strategy that unifies the two stages is the generalized additive model (GAM) framework (Hastie and Tibshirani 1990) with spatial random effects at the basic geographic unit level. A GAM allows the covariate effects to be smoothly varying instead of linear, a highly desired feature in capturing the large scale variation. The spatial random effects account for the structured, small scale variation across the basic geographic units. Additional unstructured random effects can be included to account for the spatial heterogeneity for each basic unit.

Such models are termed generalized geoadditive models (Kammann and Wand 2003) and have been applied to insurance data (Fahrmeir et al. 2003, 2007b; Lang and Brezger 2004). Inferences about generalized geoadditive models are often made in the Bayesian framework with the Markov chain Monte Carlo (MCMC) method (Fahrmeir et al. 2007a;

Figure 1. Geographic rating area map for Ohio approved by Ohio Department of Insurance in Feb. 2014. Source: http://www.insurance.ohio.gov/Company/Documents/Rating_Areas_Map.pdf.



Fahrmeir and Lang 2001; Klein et al. 2014, 2015). Due to the computing intensive MCMC, however, the model has not been widely applied to geographic rating despite its natural potential. Practical applications are further complicated by excess zeros commonly observed in claims, resulting in semi-continuous data. The most relevant work in Klein et al. (2014) used zero-inflated generalized geo-additive models in non-life ratemaking, but implementable rating areas using the spatial random effects were not suggested and actuarial implications were not fully discussed.

This paper aims to provide a practical method for geographical rating with a two-part generalized geo-additive model, using a fast alternative to MCMC, the integrated nested Laplace approximation (INLA) (Rue et al. 2009). The implementation is available in R package INLA (Martins et al. 2013), which could reduce computing time from days to hours for large data sets. Because real health insurance data

are privileged, we choose to use a simulated data to demonstrate the models and methods, which mimic the real data from a company, with features such as spatial dependence, excess number of zeros, and nonlinear covariate effects. An additional advantage of simulated data is that the true model and true parameter values are known, facilitating model fitting assessment that is otherwise not possible with real data. Our code is in supplementary materials to ensure reproducibility. The simulated data are healthcare expenses and covariates at the individual level, including the basic geographical unit where each individual resides such as zipcode or county. The expenses typically have a positive probability mass at zero (about 20% in the simulated example in Section 2), and a two-part model is used for such data (Deb et al. 2006; Frees and Sun 2010). The first part models the binary response of whether or not the expense is positive, and the second part models the expense amount if it is positive. Nonlinear covariate effects, structured spatial random effects, and unstructured random effects are included in both parts. The structured spatial random effects are in the form of Gaussian Markov random fields (GMRF) (e.g., Rue and Held 2005).

The rating for each basic geographical unit is derived as a ratio based on the two sets of structured spatial random effects. The ratings vary smoothly on the map without territorial boundaries, and they are justified by the unified fully Bayesian modeling framework. Units containing no individual data can still be rated with the built-in information borrowing mechanism of the inference procedure. The resulting ratings are more fair to customers, more accurate in predicting expenses, and more profitable for insurers. We further compare models with spatial effects at the county level and those at the zipcode level, due to regulation restrictions. Models at the county level already perform much better than models without any spatial effect, but modeling at the finer zipcode level, if allowed by regulations, could lead to additional gain. Although the methods are motivated and illustrated with healthcare insurance, they can be applied to property and casualty insurance also.

The ratings vary smoothly on the map without territorial boundaries, and they are justified by the unified fully Bayesian modeling framework. Units containing no individual data can still be rated with the built-in information borrowing mechanism of the inference procedure.

The rest of the paper is organized as follows. A simulated data set mimicking a health insurance claim data set in reality is introduced in Section 2. The generalized geoadditive model and its inferences using the INLA package are presented in Section 3. A definition of ratings at the zipcode level with the spatial random effects in the model is developed in Section 3.6. The simulated data is analyzed, with results discussed and ratings map presented in Section 4. A discussion concludes in Section 5. Details about the data generation and INLA are relegated to the Appendix.

2. Healthcare expense data

A simulated dataset is used to illustrate the proposed methods which avoids the proprietary restrictions from using real data. The data generating mechanism was designed to generate data with features mimicking those of real data that are typically available for geographic rating; see details in Appendix A. Healthcare expenses of one month for n = 20,000 members in Ohio were generated with their age, gender, income, and zipcode. The variable income does not need to be income; it can represent a continuous variable of importance other than the demographics. The use of age and gender in pricing is subject to state or federal regulations, so their usage here is for illustration purposes. The covariates of the simulated data were generated independently. In a practical setting, they can be correlated, which should be handled the same way as done in a multiple regression

Table 1. Sample rows of the simulated dataset

Expense	Zipcode	Age	Gender	(log) Income
282.84	43001	36	0	5.52
180.22	43001	33	0	7.28
347.31	43001	21	1	5.40
1523.75	43001	45	1	6.64
224.21	43001	23	1	5.42

model. Moderate correlation does not bring extra difficulty to our methodology; too high correlation causes collinearity, which can be fixed, for example, by constructing new predictors from a principal component analysis. The geographic distribution of the n members in 1197 zipcode areas and their ages were generated to approximately match those from the census data (http://www.census.gov/popest/). The resulting dataset has five variables for n members: expense, zipcode, age, gender (male = 1), and income (on log scale). Table 1 shows the first 5 rows of the dataset.

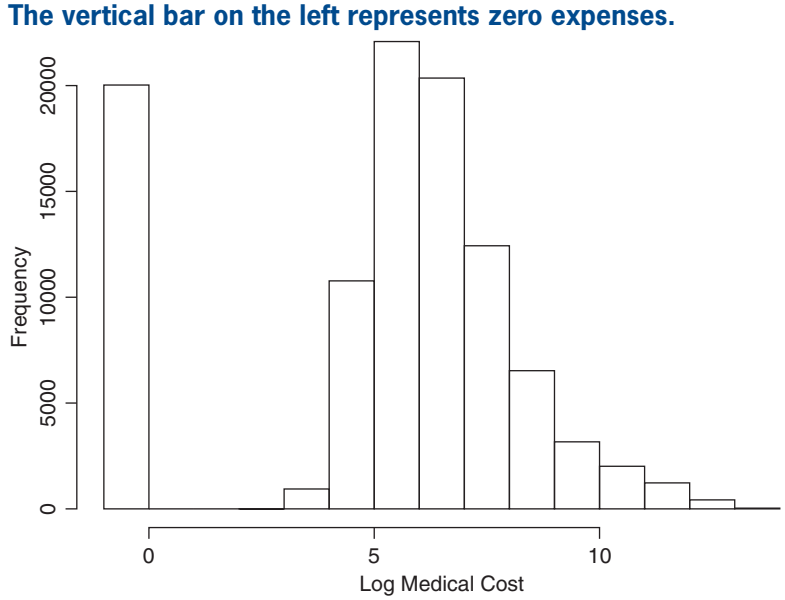
The simulated healthcare expense data has several realistic features. First, about 20% of the members have zero expense; Figure 2 shows the histogram of the healthcare expense per month on the log scale, with the leftmost bar representing the frequency of zeros. Second, children and seniors are associated with higher expenses; that is, the age effect is nonlinear, V-shaped. Lastly, everything else held constant, members in urban zipcode areas are more likely to have higher expenses than those in rural zipcode areas. This was enforced by the structural spatial effects with higher values assigned to urban zipcode areas; see details in the Appendix A. A reasonably good geographic rating method should address the challenges from these three features: 1) the positive probability of zero expense; 2) nonlinear effect for some risk factors; and 3) the smoothly varying ratings at the zipcode (or county) level. The proposed methods address these challenges by a two-part model that allows smooth covariate effects and spatial random effects. Although the spatial effects were generated at the zipcode level, it is also of interest to investigate the impact of a restriction that only allows county level geographic rating.

3. Models and methods

3.1. Generalized geoadditive model

For $i = 1, \ldots, n$, let Y_i be the response variable. In the geographic rating application, the response variable can be a binary variable indicating the

Figure 2. Histogram of healthcare expense on the log scale. The vertical bar on the left represents zero expenses.



presence/absence of healthcare expense, or the log transformed healthcare expense if it is positive. Let X_i and Z_i be a $p \times 1$ and a $q \times 1$ covariate vector, which have linear and nonlinear effects, respectively. Let s_i be the region (zipcode area) in which subject i resides, $s_i \in \{1, \ldots, R\}$, where R is the number of regions (which could be zipcode areas or counties). A generalized geoadditive model (Kammann and Wand 2003) is

$$g(\mu_i) = X_i^\top \boldsymbol{\beta} + \sum_{j=1}^{q} f_j(Z_{ij}; \boldsymbol{\alpha}) + \gamma_{s_i} + \epsilon_{s_i} \tag{3.1}$$

where g is a known link function, $\mu_i = E[Y_i \mid X_i, Z_i, \gamma_{s_i}, \epsilon_{s_i}]$, $\boldsymbol{\beta}$ is a $p \times 1$ vector of coefficients for X_i, f_j, $j = 1, \ldots, q$, are smooth nonlinear functions with parameter vector $\boldsymbol{\alpha}$, γ_{s_i} are structured spatial random effects (spatially dependent) at the region level to be further described below, and ϵ_{s_i} are unstructured random effects at the region level. The joint distribution of $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_R)^\top$ is $N(0, \tau_\epsilon^{-1} I_R)$, where I_R is identity matrix of dimension R, and τ_ϵ is a precision parameter.

Model (3.1) encompasses the GAM and GLM as special cases. For example, when γ is not present, the model is a GAM with an unstructured random effect at the region level; when neither γ nor ϵ is present, the model is a GAM; when no covariate has smooth nonlinear effect, the model is a GLM with structured and unstructured region level random effects. Since the smooth nonlinear effects f_j, $j = 1, \ldots, q$, are confounded via the intercept, constraints are needed for their identifiability. The best constraints are $\sum_{i=1}^{n} f_j(Z_{ij}) = 0$ for all j, which makes each f_j orthogonal to the intercept and leads to minimum width confidence intervals for the constrained f_j (Wood 2006).

The structured random effects γ_{s_i} in model (3.1) account for a smooth geographical surface across all the regions. In contrast to the unstructured random effects ϵ_{s_i} which capture the heterogeneity across the regions, the spatial random effects are usually surrogates of unobserved factors. As most of the unobserved factors vary smoothly over the space, two neighboring regions are more alike than two regions further apart. The local dependence structure is described by an Intrinsic GMRF, which adopts an intuitive conditional specification:

$$\gamma_s \mid \gamma_{t;t \neq s}, \tau_\gamma \sim N\left(\frac{1}{n_s}\sum_{t \sim s}\gamma_t, \frac{1}{n_s \tau_\gamma}\right), \tag{3.2}$$

where n_s is the number of neighbors of region s, $t \sim s$ indicates that region t and region s are neighbors, and τ_γ is a precision parameter. The joint distribution of $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_R)$ can be equivalently specified as

$$\boldsymbol{\gamma} \sim N\left(0, \tau_\gamma^{-1}(I_R - W)^{-1}\right), \tag{3.3}$$

where $W = (w_{st})$, $w_{st} = I(t \sim s)/n_s$. Model (3.3) specifies an improper distribution because the precision matrix $I_R - W$ is not of full rank. It can, however, be used as a prior for $\boldsymbol{\gamma}$. This model is known as an intrinsic conditional autoregressive (ICAR) model, because model (3.2) is the limit of an autoregressive model as the autoregressive coefficient $\rho = 1$. As such, it can accommodate stronger dependence than models with $\rho < 1$.

3.2. Two-Part generalized geoadditive model

As patient level healthcare expenses have a large percentage at zero, applying model (3.1) directly would not be appropriate. Two-part models have been developed for zero-modified data; see Neelon and O'Malley (2014) for a recent review. We consider a two-part generalized geoadditive model. A generalized geoadditive model is used for the probability of non-zero expense in the first part, and

We consider a two-part generalized geoadditive model. A generalized geoadditive model is used for the probability of non-zero expense in the first part, and a second generalized geoadditive model is used for the positive expense in the second part.

a second generalized geoadditive model is used for the positive expense in the second part.

Let $\mathcal{Y}_i$ be the expense of subject i, $i = 1, \ldots, n$. In the first part, the response variable is the indicator variable $Y_i^{(1)} = 1_{\mathcal{Y}_i > 0}$. By adding a superscript "(1)" to all the components in model (3.1), the first part model is

$$
\begin{aligned}
&Y_i^{(1)} | \mu_i^{(1)} \overset{ind}{\sim} \text{Bernoulli}\left(\mu_i^{(1)}\right), \qquad i = 1, \ldots, n, \\
&g^{(1)}\left(\mu_i^{(1)}\right) = X_i^\top \boldsymbol{\beta}^{(1)} + \textstyle\sum_{j=1}^{q} f_j^{(1)}\left(Z_{ij}; \boldsymbol{\alpha}^{(1)}\right) + \gamma_{s_i}^{(1)} + \epsilon_{s_i}^{(1)}, \\
&\boldsymbol{\gamma}^{(1)} \sim \text{ICAR}\left(\tau_\gamma^{(1)}\right), \\
&\boldsymbol{\epsilon}^{(1)} \sim N\left(0, \tau_\epsilon^{(1)^{-1}} I_R\right),
\end{aligned} \tag{3.4}
$$

where $\mu_i^{(1)} = E[Y_i^{(1)} | X_i, Z_i, \gamma_i^{(1)}, \epsilon_i^{(1)}]$, $\boldsymbol{\gamma}^{(1)} = (\gamma_1^{(1)}, \ldots, \gamma_R^{(1)})^\top$, $\boldsymbol{\epsilon}^{(1)} = (\epsilon_1^{(1)}, \ldots, \epsilon_R^{(1)})^\top$ and the ICAR model is defined as model (3.3). The link function $g^{(1)}$ was chosen to be the logit link in the data analysis in Section 4.

In the second part, the response variable is $Y_i^{(2)} = \log \mathcal{Y}_i$ provided $\mathcal{Y}_i > 0$. With superscript "(2)", the second part model is

$$
\begin{aligned}
&Y_i^{(2)} | \mu_i^{(2)} \overset{ind}{\sim} F\left(y; \mu_i^{(2)}, \delta\right), \qquad i = 1, \ldots, n, \\
&g^{(2)}(\mu_i) = X_i^\top \boldsymbol{\beta}^{(2)} + \textstyle\sum_{j=1}^{q} f_j^{(2)}\left(Z_{ij}; \boldsymbol{\alpha}^{(2)}\right) + \gamma_{s_i}^{(2)} + \epsilon_{s_i}^{(2)}, \\
&\boldsymbol{\gamma}^{(2)} \sim \text{ICAR}\left(\tau_\gamma^{(2)}\right), \\
&\boldsymbol{\epsilon}^{(2)} \sim N\left(0, \tau_\epsilon^{(2)^{-1}} I_R\right),
\end{aligned} \tag{3.5}
$$

where $\mu_i^{(2)} = E[Y_i^{(2)} | X_i, Z_i, \gamma_i^{(2)}, \epsilon_i^{(2)}]$, $\boldsymbol{\gamma}^{(2)} = (\gamma_1^{(2)}, \ldots, \gamma_R^{(2)})^\top$, $\boldsymbol{\epsilon}^{(2)} = (\epsilon_1^{(2)}, \ldots, \epsilon_R^{(2)})^\top$, and $F(y; \mu, \delta)$ is a distribution function specified by a mean parameter μ and possibly another parameter δ. In the data analysis in Section 4, the link function $g^{(2)}$ was the identity link and F is the normal distribution with mean μ and variance δ. Flexible distributions such as generalized gamma can be used for F in the second part to capture the skewness in the observed data (Liu et al. 2010; Manning et al. 2005). The covariates X_i and Z_i in the two parts do not have to be the same, although they are the same in the analysis in Section 4.

3.3. Bayesian inference and INLA

Models (3.4) and (3.5) contain many Gaussian components that can be taken advantage of in inference under the Bayesian framework. Consider the first part (3.4) with logit link as an example, where Gaussian prior distributions are imposed on $\boldsymbol{\beta}^{(1)\top}$ and $\boldsymbol{\alpha}^{(1)\top}$, with hyperparameter vectors $\boldsymbol{\tau}_\beta^{(1)}$ and $\boldsymbol{\tau}_\alpha^{(1)}$. The prior for $\boldsymbol{\gamma}^{(1)}$ is defined as model (3.3) with a precision parameter $\tau_{\gamma^{(1)}}$. The prior for $\boldsymbol{\epsilon}^{(1)}$ is normal with mean 0 and precision parameter $\tau_{\epsilon^{(1)}}$. Define $\boldsymbol{w}^{(1)} = (\boldsymbol{\beta}^{(1)\top}, \boldsymbol{\alpha}^{(1)\top}, \boldsymbol{\gamma}^{(1)\top}, \boldsymbol{\epsilon}^{(1)\top})^\top$, the vector of all unknown Gaussian variables of interest in the model. Let $\boldsymbol{\theta}^{(1)} = (\boldsymbol{\tau}_\beta^{(1)\top}, \boldsymbol{\tau}_\alpha^{(1)\top}, \tau_\gamma^{(1)}, \tau_\epsilon^{(1)})^\top$ and let $\pi(\boldsymbol{\theta}^{(1)})$ be the prior of $\boldsymbol{\theta}^{(1)}$. Let $\pi(\cdot|\cdot)$ denote the conditional distribution of its arguments. Then $\pi(\boldsymbol{w}^{(1)} | \boldsymbol{\theta}^{(1)})$ is Gaussian with zero mean and precision matrix $\boldsymbol{Q}(\boldsymbol{\theta}^{(1)})$.

Since Y_i are independent given all covariates and latent Gaussian variables, the posterior can be written as

$$
\begin{aligned}
&\pi\left(\boldsymbol{w}^{(1)}, \boldsymbol{\theta}^{(1)} | \boldsymbol{Y}\right) \propto \\
&\pi\left(\boldsymbol{\theta}^{(1)}\right) \pi\left(\boldsymbol{w}^{(1)} | \boldsymbol{\theta}^{(1)}\right) \textstyle\prod_{i=1}^{n} \pi\left(Y_i | w_i^{(1)}, \boldsymbol{\theta}^{(1)}\right),
\end{aligned}
$$

where $\pi(Y_i | w_i^{(1)}, \boldsymbol{\theta}^{(1)})$ is the Bernoulli likelihood under model (3.4), and $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}$. The posterior density for the second part model (3.5) can be worked out similarly with superscript "(2)".

A common approach for inference in such a model is to use MCMC. As pointed out by Rue et al. (2009), however, performance of MCMC with component-wise updates in this context is poor due to strong dependence among $\boldsymbol{w}^{(1)}$ itself, and between $\boldsymbol{w}^{(1)}$ and $\boldsymbol{\theta}^{(1)}$, especially when n and R are large. Consequently, the computational efficiency of MCMC is very low. Long chains are needed for convergence to occur, which may take days.

INLA is a new approach to statistical inference for latent Gaussian models (Martino and Rue 2010; Rue et al. 2009). It provides fast, accurate approximations to the posterior densities of parameters and latent Gaussian variables of interest. A sketch of the approximation algorithms is in Appendix B.

We refer readers to Rue et al. (2009) for more details. The INLA methodology is efficiently implemented, and includes sparse matrices operations (Martino and Rue 2009). By using INLA, days of computing time using MCMC can be reduced to hours (e.g., Carroll et al. 2015; Taylor and Diggle 2014). The software package uses OpenMP to speed up the computations for shared memory machines, i.e., multicore processors which are equipped on most personal computers nowadays. An R package is available for ease of usage, and was used in the data analysis in Section 4. The software is open-source and can be downloaded from the website www.r-inla.org, which also includes many applications and case studies.

3.4. Model comparison

Since the two parts of the model are separated using different parts of the data, model comparisons can be done for each part separately. The deviance information criterion (DIC) is a commonly used model comparison criterion in the Bayesian framework (Spiegelhalter et al. 2002). Taking the first part as an example, let $\xi^{(1)} = (\boldsymbol{\theta}^{(1)}, \boldsymbol{w}^{(1)})^\top$ be the vector of all unknown parameters of interest in the model. The DIC is defined as

$$\mathrm{DIC} = \bar{D} + p_D,$$

where $\bar{D}$, as a measure of fit, is the expectation of the deviance of the model with respect to the posterior distribution of $\xi^{(1)}$, and p_D is the effective number of parameters measuring the model complexity. Specifically, $\bar{D} = E_{\xi^{(1)}}[D(\xi^{(1)})]$ and $D(\xi^{1)}) = -2\log p(Y^{(1)}|\xi^{(1)})$. Several methods have been proposed for calculating p_D. Spiegelhalter et al. (2002) proposed to use

$$p_D = \bar{D} - D\left(\bar{\xi}^{(1)}\right),$$

where $\bar{\xi}^{(1)}$ is the posterior mean of $\xi^{(1)}$. Gelman et al. (2014a) suggested

$$p_D = \frac{1}{2}\mathrm{Var}\left[D\left(\xi^{(1)}\right)\right]. \tag{3.6}$$

INLA uses an alternative method to (3.6) proposed by Watanabe (2010),

$$p_D = \sum_{i=1}^{n} \mathrm{Var}\left[\log p\left(Y_i^{(1)}\middle|\xi^{(1)}\right)\right].$$

This method is more stable than (3.6) because it computes the variance separately for each data point and sums; the summing yields stability (Gelman et al. 2014b). A smaller DIC indicates a better model. In general, rules of thumb suggest that differences of 3 or more in DIC should be regarded as significant (Spiegelhalter et al. 2002). Although the DIC is widely used, it is known that it underpenalizes complex models with many random effects (Plummer 2008; Riebler and Held 2009).

Another popular model comparison criterion is the conditional predictive ordinate (CPO) (Geisser 1993; Pettit 1990), which is also provided in the INLA package. The CPO of observation i is the predictive density of the ith observation given the rest of the data $\boldsymbol{Y}_{-i}^{(1)}$ that excludes the ith observation:

$$\mathrm{CPO}_i = \pi\left(Y_i^{(1)}\middle|\boldsymbol{Y}_{-i}^{(1)}\right).$$

To assess the overall predictive quality of the model being considered, the logarithm of the pseudo-marginal likelihood (LPML) (Geisser and Eddy 1979) can be computed,

$$\mathrm{LPML} = \sum_{i=1}^{n} \log \mathrm{CPO}_i.$$

A scaled version –LPML/n is known as the cross-validated log-score (Gneiting and Raftery 2007). A larger value of LPML indicates a better predictive model. The difference of two models in LPML is the logarithm of the pseudo-Bayes factor (PsBF) (Dey et al. 1997; Geisser and Eddy 1979). The asymptotic distribution of PsBF (Gelfand and Dey 1994) may be used to calibrate PsBF in a similar way to calibration for Bayes factors (Kass and Raftery 1995): a difference of 1–2 indicates strong preference and more than 2 will be decisive.

3.5. Predictive performance

Not considering administrative costs, the premium of a customer with a given set of covariates is the predicted healthcare expense. The predictive performance of models can be evaluated by using hold-out data. Commonly used measures are mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean squared prediction error (RMSPE) (e.g., Frees et al. 2013). A more recent measure by Quiroz et al. (2015) brings randomness into selecting training and testing data sets. This approach can be applied to all predictive accuracy measures. Taking RMSPE as an example, the process has four steps:

1. Randomly select V out of the n observations to be the testing data and the rest to be the training data.
2. Fit proposed models on the training data.
3. Make predictions on the testing data using the posterior mean of the parameters and obtain RMSPE for each fitted model

$$\text{RMSPE} = \sqrt{\frac{1}{V}\sum_{i=1}^{V} d_i^2},$$

   where d_i is the prediction error of the ith testing data.
4. Repeat steps 1–3 M times, and take the average of the M resulting RMSPE values as the mean RMSPE (MRMSPE) for each model.

The same procedure can be applied to get mean MAE (MMAE).

The variation of each accuracy measure is also of interest. The standard deviation from the M replicates for each measure implies the stability of the performance under the measure. Smaller standard deviation is preferred.

Other measures for predictive modeling comparison such as Gini or lift score could also be used, if the goal of the prediction is ranking.

Geographic ratings can be defined based on the two-part model, where the structured spatial effects from both parts contribute.

3.6. Geographic rating

Geographic ratings can be defined based on the two-part model, where the structured spatial effects from both parts contribute. In the logistic part, the spatial random effects $\boldsymbol{\gamma}^{(1)}$ are the zipcode (or county) level adjustments on the probability of having non-zero expense for patients with the same covariate vector. The adjustments take effect on the scale of the log odds. In the lognormal part, the spatial random effects $\boldsymbol{\gamma}^{(2)}$ are the zipcode (or county) level adjustments on the log scale of the expense if the expense is positive. Ideally, the two effects should be combined to give a single adjustment for each region in geographic rating.

We propose a method motivated from the prediction of the model. Given the covariates of patient i, the mean healthcare expense is predicted to be

$$\hat{\mu}_i = \hat{\mu}_i^{(1)} \exp\left[\hat{\mu}_i^{(2)} + \hat{\delta}/2\right], \tag{3.7}$$

where $\hat{\mu}_i^{(1)}$ and $\hat{\mu}_i^{(2)}$ are estimates of $\mu_i^{(1)}$ and $\mu_i^{(2)}$ in models (3.4)–(3.5). The additional term $\hat{\delta}/2$ is due to the expectation of a log-normal random variable with variance δ on the log scale. We need to transform the individual level prediction to a regional level adjustment for ease of practical operation as needed in geographic rating.

It is clear in model (3.5) that the region level adjustment in $\hat{\mu}_i^{(2)}$ is simple due to the log link—just a scale of exp $[\hat{\gamma}_{s_i}^{(2)}]$. The region level adjustment in $\hat{\mu}_i^{(1)}$ is not readily available as seen from model (3.4): since $\hat{\mu}_i^{(1)}$ is on the scale of log odds, the impact of the spatial random effect $\gamma_{s_i}^{(1)}$ depends on the individual covariates X_i, Z_i. We propose to average over all individuals in each region to form a regional level adjustment factor on the probability of non-zero expense. For each region r, we take the mean of the linear predictor in (3.4) over all the observed individuals in this region to form a region level linear predictor. In particular, let $\bar{\eta}_r$ be the regional average

of $\hat{\eta}_{s_i} = X_{s_i}^\top \hat{\beta}^{(1)} + \sum_{j=1}^{q} \hat{f}_j^{(1)}(Z_{s_i j})$ over all i such that $s_i = r$. This regional level linear predictor $\bar{\eta}_r$ is then combined with the region level spatial random effect $\gamma_r^{(1)}$ to form a regional level probability of non-zero expense for region r

$$\hat{\phi}_r = \text{logit}^{-1}\left[\bar{\eta}_r + \hat{\gamma}_r^{(1)}\right].$$

We propose to use $\hat{\phi}_r$, $r = s_i$, as the region level adjustment in place of $\hat{\mu}_i^{(1)}$ in (3.7). Combining the two parts, the proposed region level geographic rating adjustment for region r, $r = 1, \ldots, R$, is

$$\rho_r = \hat{\phi}_r \exp\left[\hat{\gamma}_r^{(2)}\right].$$

This adjustment is a multiplicative effect in predicting the healthcare expense that is applied after the individual covariate effects. It combines effects from both parts of the two-part model. In the case where a large proportion of policyholders in a region had no expense but the remaining policyholders had very high expenses, the effect obtained from the log-normal part exp $[\hat{\gamma}_r^{(2)}]$ will be very high, but it will be downweighted by the effect from logistic part $\hat{\phi}_r$ which would be much smaller.

4. Illustration

As an illustration, the two-part model was fitted to the simulated data described in Section 2 with the logit link in the first part and the identity link on the log scale in the second part. The effect of age was set to be nonlinear in both parts. Although INLA offers completely nonparametric specifications of nonlinear effects through random walk priors, we chose to describe the nonlinear effect with basis spline regression, in which case, the coefficients of the basis are estimated the same way as those in linear effects, and the computing cost is much lower. The spline basis was constructed with the bs function in R package splines, and a sum to zero constraint was imposed on each basis such that the intercept in the model becomes identifiable. The degrees of freedom of the spline basis could be chosen with the DIC or LPML, which is beyond the scope of this manuscript. In our analysis, we used cubic splines with 5 degrees of freedom (and 2 internal knots), which provided sufficient flexibility in recovering the true curve.

The mean components of the two-part model (3.4)–(3.5) are

$$\text{logit}\left(\mu_i^{(1)}\right) = \beta_0^{(1)} + \beta_1^{(1)} G_i + \beta_2^{(1)} I_i + B_i^\top(A_i)\theta^{(1)} + \gamma_{s_i}^{(1)} + \epsilon_{s_i}^{(1)},$$

$$\mu_i^{(2)} = \beta_0^{(2)} + \beta_1^{(2)} G_i + \beta_2^{(2)} I_i + B_i^\top(A_i)\theta^{(2)} + \gamma_{s_i}^{(2)} + \epsilon_{s_i}^{(2)},$$

where G is gender, I is log income, A is age, and $B(A_i)$ represents the spline basis expansion of age. The spatial random effects $\gamma_{s_i}^{(1)}$ and $\gamma_{s_i}^{(2)}$, $s_i \in \{1, \ldots, R\}$, are possibly restricted by regulations. To investigate the impact of the regulations, we fit models with spatial effects at two different levels: zipcode level (correct specification with $R = 1197$) and county level (misspecification with $R = 88$). The models are referred to as the zipcode level rating model and the county level rating model, respectively. As the data were generated with zipcode level spatial effects, county level rating is less optimal than zipcode level rating, but may still be better than no geographic rating.

Since the smoothing effects are decomposed into several linear effects, the hyperparameters $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$ are reduced to $(\tau_\gamma^{(1)}, \tau_\epsilon^{(1)})$ and $(\tau_\gamma^{(2)}, \tau_\epsilon^{(2)}, 1/\delta)$, respectively. The priors for these hyperparameters are set on $\log\boldsymbol{\theta}^{(1)}$ and $\log\boldsymbol{\theta}^{(2)}$ in INLA as gamma distributions with shape 1 and scale 20,000. The priors for $(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$ are normal with some mean μ and precision τ which are typically unequal to the default values. Users can specify values of priors μ and τ through the inla function in the INLA package. Examples are given in the supplemental code[1]. The structured random effects $\gamma_{s_i}^{(1)}$ and $\gamma_{s_i}^{(2)}$ account for a smooth geographical surface across all the rating

[1]Available at https://elizabeth-schifano.uconn.edu/

areas while the unstructured random effects $\epsilon_{s_i}^{(1)}$ and $\epsilon_{s_i}^{(2)}$ capture the heterogeneity across the rating areas.

4.1. Statistical analysis

Posterior means, standard errors, and 95% credible intervals of the fixed effect coefficients are summarized in Table 2. With such a large sample at the individual level, these effects are all estimated quite accurately, very close to the true values (2, 0.1, 1) and (6, 1, 0.4) in the logistic part and log-normal part, respectively. The uncertainty in estimation is much higher in the logistic part than in the log-normal part, because this part contains less information due to the binary nature of the response. The effects estimated from the zipcode level rating model are very similar to those from the county level rating model, except with higher bias for the intercept and higher standard errors for the two coefficient parameters (gender and income) in the log-normal part. The estimated nonlinear effects of age are shown in Figure 3, which recover the true curves very closely in both parts of the model. Again, the uncertainty is much higher for the logistic part than for the log-normal part, and the results from models with different geographic rating levels are very similar.

The posterior mean of the spatial effects in both parts, $\gamma_s^{(1)}$ and $\gamma_s^{(2)}$, are shown in Figure 4 using 9 levels of gray scales categorized by their quantiles, with darker colors indicating higher spatial effects. For models with zipcode level and county level spatial effects, the overall color patterns of the estimates are very similar to that of the true effects in Figure 6 in Appendix A. Urban areas are estimated to have higher effects than rural areas. Nonetheless, the zipcode level spatial effects vary more smoothly over space due to its much finer resolution than the county level spatial effects. Further, the county level spatial effects have a much smaller magnitude than the zipcode level spatial effects, because the former is coarser so that zipcode effects within the same county get averaged. The finer resolution of zipcode spatial effects, if allowed by regulations, allows for improved predictions over territory rating methods based on counties or even bigger territories as shown in Figure 1.

To compare rating models with or without spatial effects, we fit several variations of the two-part model by dropping the structured spatial effects and/or the unstructured random effects at both zipcode and county levels. Model 1 has neither structured nor unstructured effects; Model 2 has zipcode level unstructured effects but no structured effects; Model 3 has zipcode level structured effects but no unstructured effects; Model 4 has both structured and unstructured effects at zipcode level; Model 5 has unstructured effects but no structured effects at county level; Model 6 has structured effects but no unstructured effects at county level; and Model 7 has both structured and unstructured effects at county level. Estimation results from Model 4 and Model 7 are reported in Table 2 and Figures 3–4. The DIC and LPML for all seven models are summarized in

Table 2. Posterior means, standard errors (SE), and 95% credible intervals (CI) of fixed effect coefficients

	Zipcode level spatial effect				County level spatial effect			
	Mean	SE	95% CI		Mean	SE	95% CI	
Logistic part								
(Intercept)	2.0155	0.0442	1.9317	2.1056	1.9716	0.0445	1.8871	2.0622
gender	0.1061	0.0403	0.0271	0.1851	0.1052	0.0402	0.0263	0.1840
income	1.0260	0.0236	0.9799	1.0724	1.0233	0.0235	0.9764	1.0685
Log-normal part								
(Intercept)	5.9993	0.0061	5.9873	6.0112	5.9228	0.0040	5.9150	5.9307
gender	0.9987	0.0017	0.9954	1.0021	0.9993	0.0043	0.9910	1.0077
income	0.4000	0.0009	0.3982	0.4017	0.3996	0.0022	0.3952	0.4040

Figure 3. Estimates of nonlinear age effects from zipcode level rating model (upper) and county level rating model (lower) in logistic regression (left) and normal regression (right) and their 95% pointwise credible intervals. The credible intervals are very tight on the right due to the large sample size. The true curves are mostly overlapped by the estimated curve.

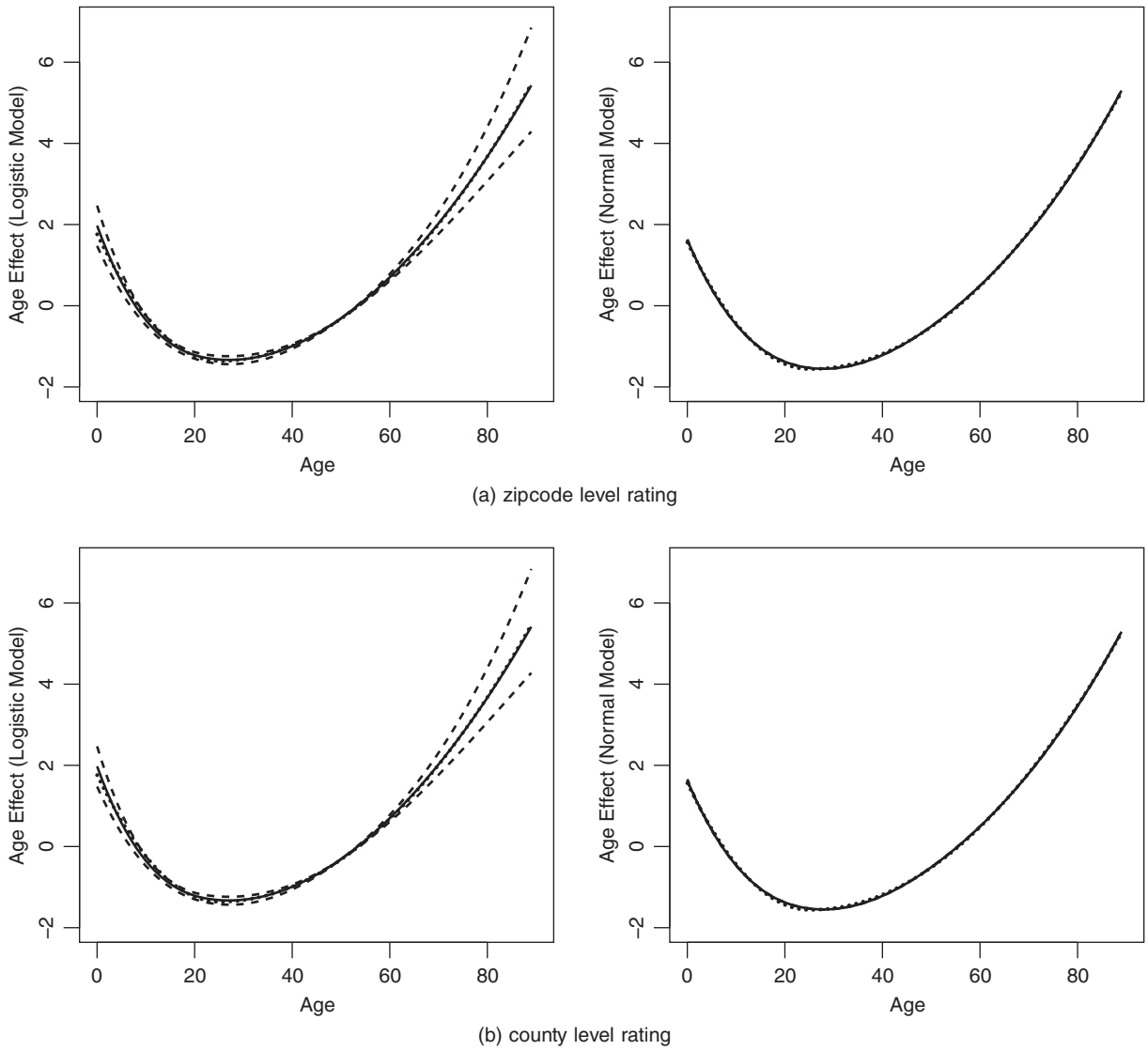
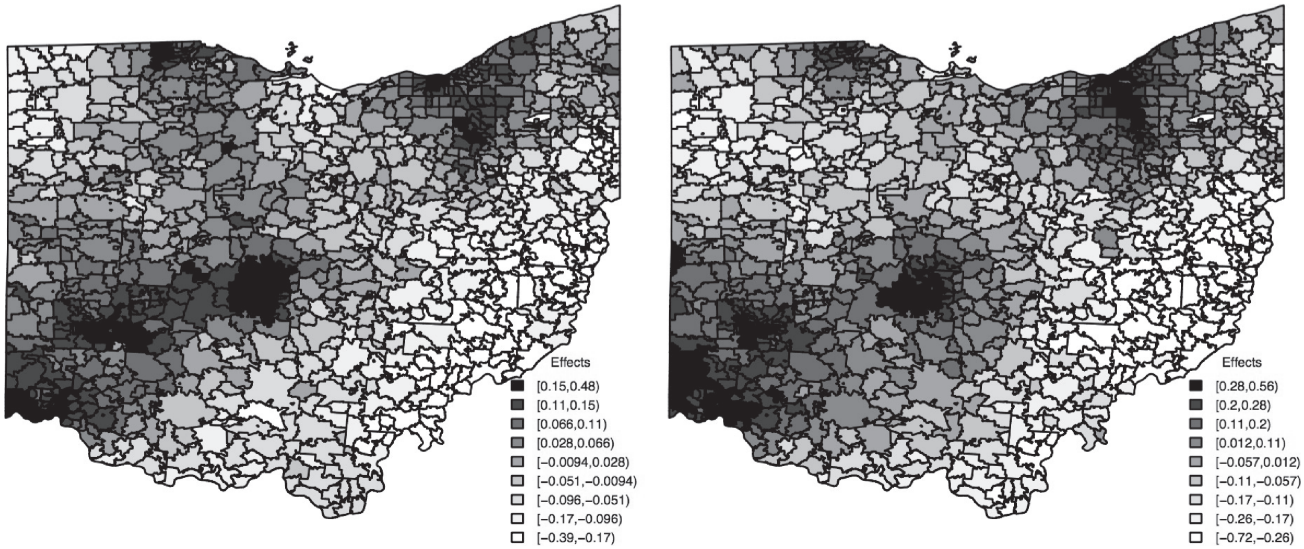


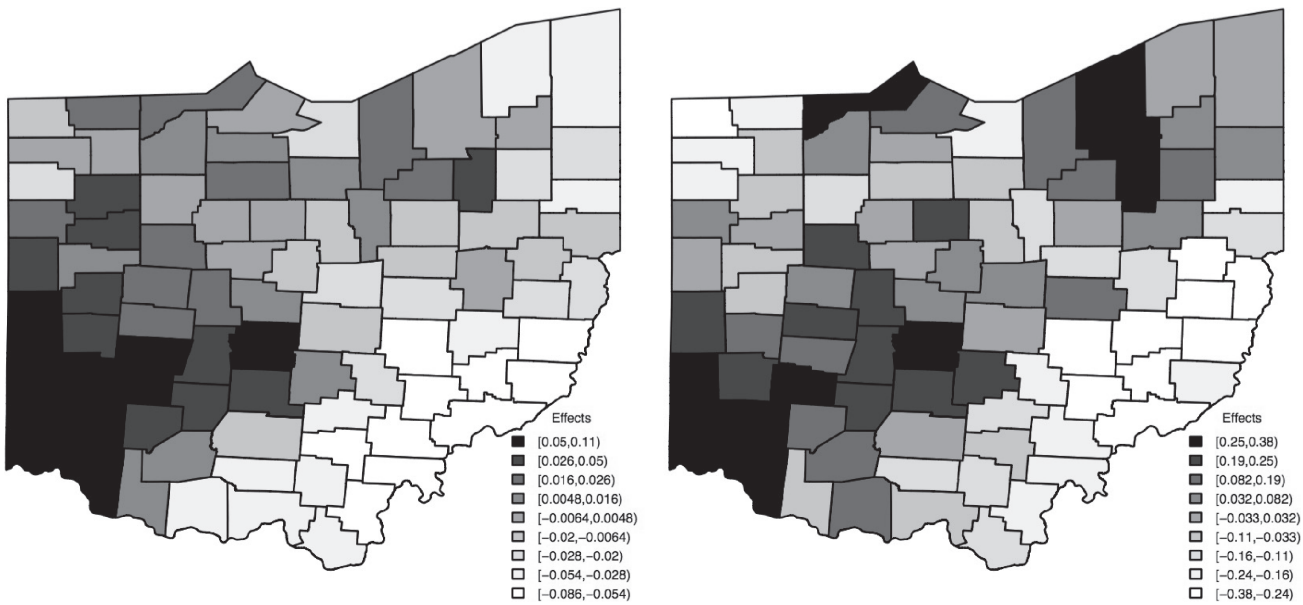
Table 3. Some numbers appear identical in the table only because of rounding. Model 1, with no spatial component, is clearly out-performed by all others. Among the zipcode level rating models, the correctly specified Model 4 is the winner in both parts, but its advantage is much clearer in the log-normal part than in the logistic part in both DIC and LPML. The county level rating models are quite competitive in the logistic part, but even their best competitor, Model 7, is far inferior to Model 4 in the log-normal part in both DIC and LPML. The comparison suggests that, if the spatial effects in the data were at a finer scale like zipcode, which is often a reasonable assumption, then modeling it at a larger scale due to regulation restrictions is not ideal but can be much better than not modeling the spatial effects at all.

The models we considered are only a few choices from many possibilities. Other models using a Tweedie distribution combined with spatial random effects could be competitive for a real data where the true distribution is unknown. The models we presented are for illustration purposes, and the model

Figure 4. Maps of structured spatial effects in zipcode level rating model (upper) and county level rating model (lower). Structured spatial effects for probability of non-zero expense (left) and expense given non-zero expense (right).



(a) zipcode level rating



(b) county level rating

Table 3. Model fitting comparison results. Model 1 has neither structured nor unstructured effects; Model 2 and Model 5 have unstructured effects but no structured effects; Model 3 has Model 6 have structured effects but no unstructured effects; Model 4 and Model 7 have both structured and unstructured effects. Models 2–4 are zipcode level rating; Models 5–7 are county level rating.

			Zipcode level rating			County level rating		
	Criteria	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Logistic	DIC	15443.3	15443.3	15437.4	15436.7	15443.1	15434.9	15436.4
	LPML	–7721.7	–7721.7	–7719.8	–7719.4	–7721.6	–7717.6	–7718.3
Log-normal	DIC	49426.2	–25912.4	–25921.9	–25937.6	–1945.5	–1941.3	–1947.8
	LPML	–24713.1	12820.8	12852.1	12869.8	974.0	969.2	975.0

with the correct specification is expected to perform the best given the data size.

4.2. Actuarial implications

To compare the predictive performance of the models in pricing, we use MMAE and MRMSPE computed with V = 5,000 and M = 100. That is, instead of fitting the model to all 20,000 observations, at each replicate, we randomly split the data into a training data of 15,000 observations and a testing data of 5,000 observations. MAE and RMSPE are obtained on the testing data based on the fit to the training data for each model given each replicate of training/testing division. MAPE is not used because the true expenses could be exactly zero.

Table 4 summarizes the MMAE and MRMSPE from 100 replicates. Model 1 which has no geographic rating component has the largest prediction errors, the zipcode level rating models (Models 2–4) have much smaller errors, and the county level rating models (Models 5–7) fall in between. The standard deviations of both measures have the same pattern. Within the zipcode level rating models, the correctly specified model (Model 4) does better, but the differences are relatively small. The three county level rating models have very close errors. These errors and their standard deviations have units in dollars. They reflect the accuracy of pricing based on the model given covariates. The results suggest that rating models with spatial effects do improve the predictive power, and if modeling them at the zipcode level is not allowed or possible, modeling them at the coarser, county level is still well worth it. Moreover, note that around 200 out of 1197 zipcode areas have no observations in the training data at each replicate, but the zipcode level rating model can still estimate the spatial effect of these areas by using information from their neighboring areas.

Table 4. Predictive performance comparison of 7 competing models. Model 1 has neither structured nor unstructured effects; Model 2 and Model 5 have unstructured effects but no structured effects; Model 3 has Model 6 have structured effects but no unstructured effects; Model 4 and Model 7 have both structured and unstructured effects. Models 2–4 are zipcode level rating; Models 5–7 are county level rating. Results are obtained by averaging over 100 replicates where the data are resampled at each replicate. Standard deviations are presented in parentheses.

Model	MMAE (SD)	MRMSPE (SD)
No geographic rating		
Model 1	1866.7 (134.6)	11416.1 (1779.3)
Zipcode level rating		
Model 2	468.5 (34.9)	2741.7 (513.3)
Model 3	454.8 (31.0)	2521.9 (416.6)
Model 4	453.3 (30.7)	2505.7 (420.0)
County level rating		
Model 5	1016.9 (84.6)	6790.7 (1458.8)
Model 6	1016.9 (84.8)	6791.8 (1461.4)
Model 7	1016.7 (84.6)	6789.2 (1458.4)

Since healthcare expenses have a probability mass at zero and are highly skewed to the right, the predictions on zeros and extremely large expenses could have very different performance across different models. Thus, it may also be valuable to compare the predictive power within each observation group categorized by the values of the expenses. Eight brackets are created based on the percentiles of the expenses. The first bracket are all zero which comprises of about 20% of the data, and the last bracket contains expenses over $2,000, which also comprises of about 20% of the data on the right tail. Table 5 summarizes the MRMSPE of Models 1, 4, and 7, and their standard deviations from 100 replicates. In bracket 1, Model 7 gives the lowest MRMSPE with a small edge over Model 4, but in all other brackets, Model 4 is the clear winner with much reduced MRMSPE and standard deviation. Both Model 4 and 7 do much better than Model 1. In particular, in the last bracket with large expenses, which is of most concern to insurers, the MRMSPE of Model 4 is about less than a half of that of Model 7, and less than a quarter of that of Model 1.

Finally, we plot the maps of geographic ratings ρ_r's proposed in Section 3.6 for Ohio in Figure 5, from the zipcode level rating model (left) and the county level rating model (right). As expected, the maps have patterns very similar to those of the structured spatial effect maps in Figure 4 because the estimated

Table 5. MRMSPE comparison results on groups of healthcare expense observations. Model 1 has neither structured nor unstructured effects; Model 4 has both structured and unstructured effects at zipcode level. Model 7 has both structured and unstructured effects at county level; Predictions are divided into 8 brackets based on the percentiles of the true responses where the first and the last brackets each contains around 20% of data. Standard deviations are presented in parentheses under each mean.

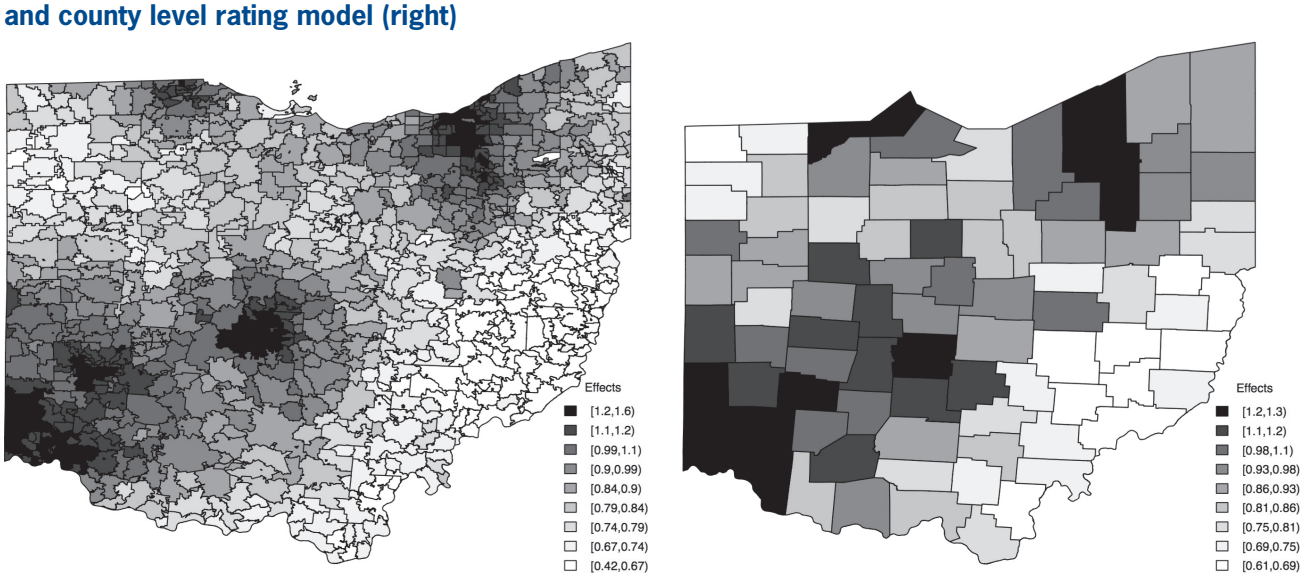
Brackets	[0, 0]	(0, 200]	(200, 400]	(400, 600]	(600, 800]	(800, 1000]	(1000, 2000]	(2000, + ∞]
Model 1	1196.1	87.6	187.8	291.1	431.7	591.0	858.4	26408.7
(SD)	(971.5)	(3.1)	(6.8)	(13.7)	(26.3)	(45.3)	(53.9)	(4138.9)
Model 4	1229.8	42.0	74.8	104.0	122.2	143.3	193.1	5600.7
(SD)	(848.8)	(0.9)	(1.9)	(3.8)	(4.6)	(6.7)	(7.3)	(918.6)
Model 7	1095.9	45.2	96.3	149.3	209.7	263.9	411.6	15667.1
(SD)	(787.1)	(1.3)	(2.3)	(4.6)	(8.3)	(11.4)	(17.4)	(3394.4)

probabilities of non-zero expense at all zipcode areas or counties have a small range. The two maps in Figure 5 are similar in darkness pattern, but their scales are very different; the range is (0.42,1.6) from the zipcode level rating and (0.61,1.30) from the county level rating. This means the former has much more flexibility than the latter to adjust the premium to account for the spatial variation that cannot be explained by the fixed effects such as gender, age, and income. Compared to the map in Figure 1 which has solid boundaries, the zipcode level rating in the left panel of Figure 5 changes in a much smoother way over all the zipcode areas on the whole map. The smoothing borrows strength from neighboring zipcode areas locally, and eliminates the need for justifying boundary lines in the traditional method.

5. Discussion

We proposed a new geographic rating method based on a two-part model with a spatial random effect in each part of the model. The method removes the solid boundaries between traditional rating areas and regards each basic geographical region such as zipcode area as an individual rating area of its own. The structured spatial random effects enforce smoothness on the resulting ratings at the basic region level. The method can be carried out with the computationally

Figure 5. Geographic ratings for Ohio, estimated from the zipcode level rating model (left) and county level rating model (right)



efficient INLA package in a Bayesian framework without the standard MCMC. The ratings can be modified frequently upon the arrival of new data. As the two-part geoadditive model gives more precise predictions of the expenses of interest, the method has the potential of substantial profit gains for insurance companies if adopted. Even when the zipcode level rating is restricted by regulations, modeling the spatial effects on a coarser spatial scale (e.g., county level) still gives better rating than not modeling them in both the statistical and actuarial sense. The methods can be applied to property and casualty insurance, in particular personal markets which may have more flexibility in geographic rating.

In real application, some practical issues will need to be considered. For example, some healthcare expenses due to inpatients or emergency room usage may be too extreme to be observed from lognormal distributions. Extreme value analysis may help to address the events at the tail. A three-part model can be constructed with a logistic model for zero or nonzero expense, a log-normal model for moderate-sized expenses, and an additional extreme model for overly large expenses. Another issue is how the INLA package can handle big data. If the method is applied to data from multiple states, the number of policies, as well as the number of zipcode area, increases significantly. The dimension of the neighborhood structure matrix increases very fast which may cause some difficulties in computation, so application under the big spatial data would be challenging. The two-part model with spatial random effects is a general modeling framework, which does not restrict specific distributional assumptions to the ones illustrated in our demonstration. The lognormal distribution of the nonzero expenses could be replaced with gamma distributions (the computational advantage from INLA will be limited by the distributions implemented in it, though). The compound Poisson distribution or Tweedie distribution may be combined with spatial random effects in principle. When the individual data has uneven exposure in practice, the exposure can often be included as an offset in the model.

Acknowledgment

The authors thank Drs. Emiliano Valdez and Jiafeng Sun for stimulating discussions on the actuarial implications of the proposed methods.

References

Besag, J., J. York, and A. Mollié, "Bayesian Image Restoration with Two Applications in Spatial Statistics," *Annals of the Institute of Statistical Mathematics* 43 (1), 1991, pp. 1–20.

Brubaker, R. E., "Geographic Rating of Individual Risk Transfer Costs without Territorial Boundaries," Casualty Actuarial Society *Forum*, 1996, pp. 97–127.

Carroll, R., A. Lawson, C. Faes, R. Kirby, M. Aregay, and K. Watjou, "Comparing INLA and OpenBUGS for Hierarchical Poisson Modeling in Disease Mapping," *Spatial and Spatio-Temporal Epidemiology* 14, 2015, pp. 45–54.

Christopherson, S., and D. Werland, "Using a Geographic Information System to Identify Territory Boundaries," in *Casualty Actuarial Society Forum*, 1996, pp. 191–211.

Deb, P., M. K. Munkin, and P. K. Trivedi, "Bayesian Analysis of the Two-Part Model with Endogeneity: Application to Healthcare Expenditure," *Journal of Applied Econometrics* 21 (7), 2006, pp. 1081–1099.

Dey, D. K., M.-H. Chen, and H. Chang, "Bayesian Approach for Nonlinear Random Effects Models," *Biometrics* 53, 1997, pp. 1239–1252.

Fahrmeir, L., T. Kneib, and S. Lang, *Regression*, New York: Springer, 2007a.

Fahrmeir, L., and S. Lang, "Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors," *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 50 (2), 2001, pp. 201–220.

Fahrmeir, L., S. Lang, and F. Spies . Generalized geoadditive models for insurance claims data. *Blätter der DGVFM* 26 (1), 2003, pp. 7–23.

Fahrmeir, L., F. Sagerer, and G. Sussmann, "Geoadditive Regression for Analyzing Small-Scale Geographical Variability in Car Insurance," *Blätter der DGVFM* 28 (1), 2007b, pp. 47–65.

Frees, E. W., X. Jin, and X. Lin, "Actuarial Applications of Multivariate Two-Part Regression Models," *Annals of Actuarial Science* 7 (2), 2013, pp. 258–287.

Frees, E. W., and Y. Sun, "Household Life Insurance Demand: A Multivariate Two-Part Model," *North American Actuarial Journal* 14 (3), 2010, pp. 338–354.

Geisser, S., *Predictive Inference*, vol. 55, Boca Raton, FL: CRC Press, 1993.

Geisser, S., and W. F. Eddy, "A Predictive Approach to Model Selection," *Journal of the American Statistical Association* 74 (365), 1079, pp. 153–160.

Gelfand, A. E., and D. K. Dey, "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society. Series B (Methodological)* 56, 1994, pp. 501–514.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, vol. 2. Taylor & Francis, 2014a.

Gelman, A., J. Hwang, and A. Vehtari, "Understanding Predictive Information Criteria for Bayesian Models," *Statistics and Computing* 24 (6), 2014b, pp. 997–1016.

Gneiting, T., and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association 102* (477), 2007, pp. 359–378.

Grize, Y. L., "Applications of Statistics in the Field of General Insurance: An Overview," *International Statistical Review* 83, 2015, pp. 135–159.

Haberman, S., and A. E. Renshaw, "Generalized Linear Models and Actuarial Science," *Statistician* 45, 1996, pp. 407–436.

Hastie, T. J., and R. J. Tibshirani, *Generalized Additive Models*, vol. 43. Boca Raton, FL: CRC Press, 1990.

Jennings, P. J., "Using Cluster Analysis to Define Geographical Rating Territories," in *2008 CAS Discussion Paper Program: Applying Multivariate Statistical Models*, pp. 34–52. Casualty Actuarial Society, 2008.

Kammann, E. E., and M. P. Wand, "Geoadditive Models," *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 52 (1), 2003 pp. 1–18.

Kass, R. E., and A. E. Raftery, "Bayes Factors," *Journal of the American Statistical Association* 90 (430), 1995, pp. 773–795.

Klein, N., M. Denuit, S. Lang, and T. Kneib, "Nonlife Ratemaking and Risk Management with Bayesian Generalized Additive Models for Location, Scale, and Shape," *Insurance: Mathematics and Economics* 55, 2014, pp. 225–249.

Klein, N., T. Kneib, and S. Lang, "Bayesian Generalized Additive Models for Location, Scale, and Shape for Zero-Inflated and Overdispersed Count Data," *Journal of the American Statistical Association* 110, (509), 2015, pp. 405–419.

Kofman, M. and K. Pollitz, *Health Insurance Regulation by States and the Federal Government: A Review of Current Approaches and Proposals for Change*, technical report, Washington, DC: Georgetown University, Health Policy Institute, 2006.

Lang, S., and A. Brezger, "Bayesian P-splines," *Journal of Computational and Graphical Statistics* 13 (1), 2004, 183–212.

Liu, L., R. L. Strawderman, M. E. Cowen, and Y.-C. T. Shih, "A Flexible Two-Part Random Effects Model for Correlated Medical Costs," *Journal of Health Economics* 29 (1), 2010, pp. 110–123.

Manning, W. G., A. Basu, and J. Mullahy, "Generalized Modeling Approaches to Risk Adjustment of Skewed Outcomes Data," *Journal of Health Economics* 24 (3), 2005, 465–488.

Martino, S., and H. Rue, *Implementing Approximate Bayesian Inference Using Integrated Nested Laplace Approximation: A Manual for the INLA Program*. Department of Mathematical Sciences, NTNU, Norway. 2009.

Martino, S. and H. Rue, "Case Studies in Bayesian Computation Using INLA," in P. Mantovan and P. Secchi (eds.), *Complex Data Modeling and Computationally Intensive Statistical Methods*, pp. 99–114. Springer. 2010.

Martins, T. G., D. Simpson, F. Lindgren, and H. Rue, "Bayesian Computing with INLA: New Features," *Computational Statistics and Data Analysis* 67, 2013, pp. 68–83.

McClenahan, C. L., "Ratemaking," in J. L. Teugels and B. Sundt (eds.), *Encyclopedia of Actuarial Science*. Wiley Online Library, 1990.

McCullagh, P., "Generalized Linear Models," *European Journal of Operational Research* 16 (3), 1984, pp. 285–292.

Miller, M. J., "Determination of Geographical Territories," presentation given at CAS Ratemaking Seminar, 2004.

NAIC and CIPR, *Health Insurance Rate Regulation*. Technical report, National Association of Insurance Commissioners and Center for Insurance Policy & Research, 2011.

Neelon, B., and A. J. O'Malley, "Two-Part Models for Zero-Modified Count and Semicontinuous Data," in B. Sobolev and C. Gatsonis (eds.), *Handbook of Health Services Research*. Springer. Forthcoming. 2014.

Pettit, L. I., "The Conditional Predictive Ordinate for the Normal Distribution," *Journal of the Royal Statistical Society, Series B (Methodological)* 52, 1990, pp. 175–184.

Plummer, M., "Penalized Loss Functions for Bayesian Model Comparison," *Biostatistics* 9 (3), 2008, pp. 523–539.

Quiroz, Z. C., M. O. Prates, and H. Rue, "A Bayesian Approach to Estimate the Biomass of Anchovies off the Coast of Peru," *Biometrics* 71 (1), 2015, pp. 208–217.

Riebler, A., and L. Held, "The Analysis of Heterogeneous Time Trends in Multivariate Age–Period–Cohort Models," *Biostatistics* 11, 2009, pp. 57–59.

Rue, H., and L. Held, *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton, FL: CRC Press, 2005.

Rue, H., S. Martino, and N. Chopin, "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71 (2), 2009, pp. 319–392.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde, "Bayesian Measures of Model Complexity and Fit," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64 (4), 2002, pp. 583–639.

Taylor, B. M., and P. J. Diggle, "INLA or MCMC? A Tutorial and Comparative Evaluation for Spatial Prediction in Log-Gaussian Cox Processes," *Journal of Statistical Computation and Simulation* 84 (10), 2014, pp. 2266–2284.

Watanabe, S., "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory," *Journal of Machine Learning Research* 11, 2010, pp. 3571–3594.

Weibel, E. J. and J. P. Walsh, "Territory Analysis with Mixed Models and Clustering," in *2008 CAS Discussion Paper Program: Applying Multivariate Statistical Models*, pp. 91–169, Arlington, VA: Casualty Actuarial Society, 2008.

Werner, G., and C. Modlin, *Basic Ratemaking,* 4 ed., Arlington, VA: Casualty Actuarial Society, 2010.

Wood, S. N., *Generalized Additive Models: An Introduction with R*, New York: Chapman and Hall/CRC, 2006.

Yao, J., "Clustering in Ratemaking: Applications in Territories Clustering," in *2008 CAS Discussion Paper Program: Applying Multivariate Statistical Models*, pp. 170–192. Arlington, VA: Casualty Actuarial Society, 2008.

Appendices

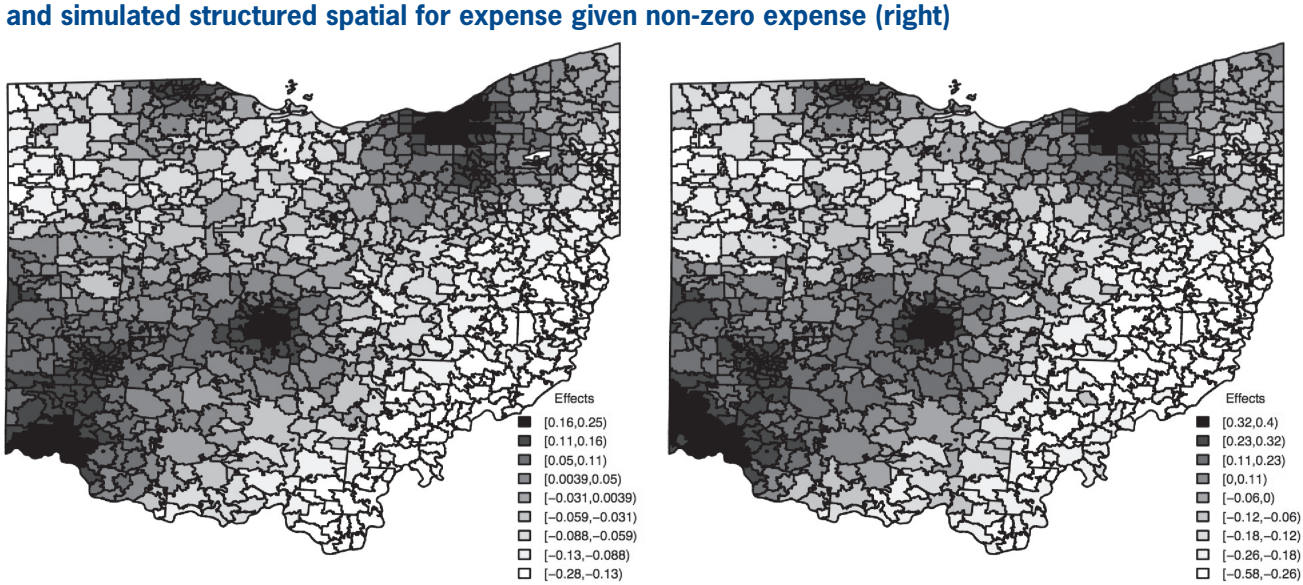
A. Generation of the simulated data

To generate from model (3.4) and model (3.5), we need to generate variables income (I), gender (G), age (A), geographic variable zipcode, corresponding structured spatial effects $\boldsymbol{\gamma}^{(1)}$ and $\boldsymbol{\gamma}^{(2)}$, and unstructured effects $\boldsymbol{\epsilon}^{(1)}$ and $\boldsymbol{\epsilon}^{(2)}$. Probability p and healthcare expenses $\mathcal{Y}$ will be generated through certain link functions and transformations.

The whole procedure is shown in steps as follows:

1. Simulate linear and non-linear smooth effects:
   - Generate variable household Income (I) in ten thousands under log scale from $N(6, 1)$.
   - Generate variable Gender (G) from a Bernoulli distribution with probability 0.5.
   - Based on demographic information from the census, generate variable Age (A) from the histogram of true distribution. A file of population distribution by age has been attached in the supplemental material (available at https://elizabeth-schifano.uconn.edu/).
   - Non-linear effect Age of normal part has quadratic form $(A - 25)^2/200$ for $A < 25$ and $(A - 25)^2/600$ for $A \geq 25$. Effect of logistic part has the same curvature in a ten times smaller scale. Both sets of non-linear effects have been centered.
2. Simulate zipcodes and structured and unstructured geospatial effects:
   - Generate geographical variable zipcode from 1197 zipcodes of Ohio. Weights are obtained from the population distribution over zipcode areas from the census. A file of such distribution has been attached in the supplemental material.
   - Generate structured spatial effects $\boldsymbol{\gamma}_s^{(1)}$ and $\boldsymbol{\gamma}_s^{(2)}$ from the ICAR model, conditioning on a set of predetermined positive spatial effects for the largest six cities in Ohio (CrICAR function in the companion code). Two sets of effects

Figure 6. Simulated structured spatial effects for probability of non-zero expense (left) and simulated structured spatial for expense given non-zero expense (right)



are simulated, one for the logistic part with conditional effects (0.375, 0.35, 0.325, 0.3, 0.275, 0.25) and precision parameter 10 (standard deviation 0.1) and one for the normal part with conditional effects (0.375, 0.35, 0.325, 0.3, 0.275, 0.25)/2 and precision parameter 5 (standard deviation 0.2). Simulated structured effects for both parts are plotted in Figure 6 which can be compared with fitted structured effects in Figure 4.

- Generate independent unstructured random effects for all zipcodes $\boldsymbol{\epsilon}^{(1)}$ from $N(0, 0.15^2)$ for the logistic part, and $\boldsymbol{\epsilon}^{(2)}$ from $N(0, 0.2^2)$ for the lognormal part.

3. Simulate binary responses for the logistic part and healthcare expenses under log scale for non-zero logistic responses.
   - Generate indicators of zero or non-zero expense. Generate η as a combination of linear effects income and gender, non-linear effect age, structured spatial effects and unstructured random spatial effects. Apply inverse logit transformation to η to get p and generate from Bernoulli distribution with probability p. After tuning parameters, the percentage of non-zero expenses is 79.98%. The true model is

$$\text{logit}[\Pr(\mathcal{Y}_i > 0)] = 2 + 1I_i + 0.1G_i + f^{(1)}(A_i) + \gamma^{(1)}_{s_i} + \epsilon^{(1)}_{s_i},$$

   - Generate patient level healthcare expenses under log scale for those with non-zero indicator values as a combination of linear effects income and gender, non-linear effect age, structured spatial effects, unstructured random effects and a random noise from $N(0,0.15^2)$. Finally, take exponential to make the expenses realistic. The true model is

$$\log \mathcal{Y}_i = 6 + 0.4I_i + G_i + f^{(2)}(A_i) + \gamma^{(2)}_{s_i} + \epsilon^{(2)}_{s_i}.$$

The R script in the supplemental material helps to replicate the case study in this paper. One needs to feed the genSP function a list of positive effects for large cities, a correlation coefficient among all regions and a value for the precision parameter to simulate spatial effects. To generate data, the number of observations (n), standard deviations for both sets of random effects ($\boldsymbol{\gamma}^{(1)}$, $\boldsymbol{\gamma}^{(2)}$) and random noise ($\boldsymbol{\epsilon}^{(1)}$, $\boldsymbol{\epsilon}^{(2)}$) need to be specified, as well as two sets of linear coefficients ($\boldsymbol{\beta}^{(1)}$, $\boldsymbol{\beta}^{(2)}$). After tuning parameters, the percentage of non-zero expenses is 79.98% and the average non-zero expense is $5,350. All values of these parameters used in this paper are listed in the R script.

B. Bayesian inference and INLA

Latent Gaussian models are hierarchical models which assume a d-dimensional Gaussian field $\boldsymbol{w}$ to be point-wise observed through n conditional independent data $\boldsymbol{Y}$ (Martino and Rue 2009). Both the covariance matrix of the Gaussian field $\boldsymbol{w}$ and the likelihood model $y_i|\boldsymbol{w}$ can be controlled by some unknown hyper-parameters $\boldsymbol{\theta}$. The posterior then reads:

$$\pi(\boldsymbol{w}, \boldsymbol{\theta}|Y) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{w}|\boldsymbol{\theta})\prod_{i=1}^{n}\pi(Y_i|w_i, \boldsymbol{\theta}),$$

where π denotes the probability density function, and Y_i's are independent conditional on w_i and $\boldsymbol{\theta}$. Generalized geoadditive is one specification of latent Gaussian models.

Integrated Nested Laplace Approximation (INLA) is a new approach to statistical inference for latent Gaussian models (Martino and Rue 2010; Rue et al. 2009). The posterior marginal distributions of interest from the latent Gaussian model can be written as

$$\pi(w_i|\boldsymbol{Y}) = \int \pi(w_i|\boldsymbol{\theta}, \boldsymbol{Y})\pi(\boldsymbol{\theta}|\boldsymbol{Y})d\boldsymbol{\theta},$$

$$\pi(\theta_j|\boldsymbol{Y}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{Y})d\boldsymbol{\theta}_{-j},$$

where $\boldsymbol{\theta}_{-j}$ stands for a vector of unknown hyper-parameters excluding the jth element.

The INLA approach uses these marginals to construct nested approximations

$$\tilde{\pi}(w_i|\boldsymbol{Y}) = \int \tilde{\pi}(w_i|\boldsymbol{\theta}, \boldsymbol{Y})\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})d\boldsymbol{\theta},$$

$$\tilde{\pi}(\theta_j|\boldsymbol{Y}) = \int \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})d\boldsymbol{\theta}_{-j},$$

where $\tilde{\pi}(\cdot|\cdot)$ is an approximated density. Thus, the approximations eliminate the need for MCMC. INLA provides accurate approximations to the marginal posterior density for the hyper-parameters $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})$ and for the full conditional posterior marginal densities for the latent variables $\tilde{\pi}(w_i|\boldsymbol{\theta}, \boldsymbol{Y})$. The approximation is based on the Laplace approximation, and for $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$ three different approaches are possible: Gaussian, full Laplace, and simplified Laplace (Rue et al. 2009). Each of these has different features, computing times, and accuracy. The Gaussian approximation is the fastest to compute but there can be errors in the location of the posterior mean or errors due to lack of skewness. The Laplace approximation is the most accurate but its computation can be time-consuming. Rue et al. (2009) suggested the simplified Laplace approximation as a compromise, which is fast to compute and usually accurate enough.

Posterior marginals for the latent variables $\tilde{\pi}(w_i|\boldsymbol{Y})$ are computed via numerical integration as:

$$\tilde{\pi}(w_i|\boldsymbol{Y}) = \int \tilde{\pi}(w_i|\boldsymbol{\theta}, \boldsymbol{Y})\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})d\boldsymbol{\theta}$$

$$\approx \sum_{k=1}^{K} \tilde{\pi}(w_i|\theta_k, \boldsymbol{Y})\tilde{\pi}(\theta_k|\boldsymbol{Y})\Delta_k.$$

The sum is over values of θ with area weights Δ_k. Posterior marginals for the hyper-parameters $\tilde{\pi}(\theta_j|\boldsymbol{Y})$ can be computed in a similar way. More information, theories, and practicalities are discussed in Rue et al. (2009).

In INLA, the prior for $\boldsymbol{\gamma}$ is ICAR with precision parameter τ_γ (Besag et al. 1991). To ensure identifiability of the overall level, a sum-to-zero constraint must be imposed on $\boldsymbol{\gamma}$. This model is specified with model = "besag" in INLA. The prior for the precision parameter τ_γ is represented as a gamma distribution on $\log\tau_\gamma$ with shape $a = 1$ and rate $b = 0.01$ by default (Martino and Rue 2010). These priors influence how smooth the spatial effects can be.

More details on implementation are available in the INLA manual (Martino and Rue 2009).