Uma aplicação de modelos latentes gaussianos em prêmios de seguro de automóveis

Marcel Irving Pereira Melo [1], **Reinaldo Antonio Gomes Marques** [2], **William Moreira Lima Neto** [3] **Júlio Sílvio de Sousa Bueno Filho** [4]

1 Introdução

Modelos para precificação do prêmio de seguros de automóveis tem sido propostos e a ferramenta estatística mais comum para este tipo de análise de dados foi por muitos anos a de modelos lineares generalizados (MLG) (ANTONIO; BEIRLANT, 2007). O grande mérito dessa abordagem é a generalização da variável resposta, que pode ser estendida para distribuições que pertençam à família exponencial. No contexto atuarial o uso do MLG se deve ao fato de que existe uma vasta aplicação dos modelos da família exponencial na análise de risco, como por exemplo, a modelagem da frequência de sinistros através das distribuições Binomial, Poisson, Binomial Negativa, e da severidade, através das distribuições Gama e Normal Inversa ((DICKSON; 2010); . Todavia, a análise por modelos lineares generalizados requer a suposição de uma amostra de variáveis aleatórias independentes, o que, geralmente, não é satisfeito para dados de seguros.

Os modelos lineares generalizados mistos (MLGM), que são extensões dos modelos lineares generalizados com a inclusão de efeitos aleatórios, surgem como alternativa quando não há independência, pois determinam a estrutura de correlação entre as observações e podem modelar dados binários, de contagem, assim como dados agrupados e dados longitudinais (McCULLOCH et al.; 2008).

Uma forma de modelo mais geral pode ser ainda definida. Os modelos latentes gaussianos (LGM) são uma subclasse de modelos de regressão de estrutura aditiva (FAHRMEIR; TUTZ, 2013) e englobam uma gama de modelos bayesianos hierárquicos, tais como: os modelos lineares generalizados (mistos), modelos com efeitos espaciais, temporais, espaço-temporais e muitos outros, de maneira que tais modelos permitem uma vasta quantidade de aplicações (RUE; HELD, 2005).

2 Materiais e Métodos

2.1 Conjunto de Dados

Os dados são referentes aos valores de prêmio de seguro de automóveis entre os anos de 2007 a 2014 por unidade da federação (UF) e foram obtidos junto ao AUTOSEG - Sistema de Estatísticas de Automóveis da Superintendência de Seguros Privados. O banco de dados contém ao todo 2160 observações e é constituído pelas variáveis: prêmio

[1]Departamento de Estatística - Universidade Federal de Lavras, Lavras, MG. e-mail: *mipmelo90@gmail.com*

[2]Laboratório de Risco Atuarial e Inovação- Universidade Federal de Alfenas, Varginha, MG. e-mail: *reinaldo.marques@unifal-mg.edu.br*

[3]Instituto de Matemática e Estatística - Universidade do Estado do Rio de Janeiro, Rio de Janeiro, RJ. em-mail: *wmlneto@gmail.com*

[4]Departamento de Estatística - Universidade Federal de Lavras, Lavras, MG e-mail: *jssbueno@des.ufla.br*

médio (PM), sexo do condutor, faixa etária do condutor e importância segurada média (IS).

2.2 Modelo para o prêmio médio de seguros

Para a análise, será descrito um modelo geral que servirá de base para todos os modelos ajustados neste estudo. O modelo geral, que possui diversas aplicações em epidemiologia, e é similar ao modelo descrito em Bernardinelli et al. (1995), que ainda sugere que efeitos de área e tendência temporal sejam tratados como aleatórios. O modelo para a aplicação de seguro é apresentado a seguir:

$$PM_i|\mu_i, \phi \sim \Gamma\left(\phi, \frac{\phi}{\mu_i}\right)$$

$$\eta_i = \log(\mu_i) = \log(IS_i) + \beta_0 + \boldsymbol{z}_i^\top \boldsymbol{\beta} + u_k + v_k + \gamma_t + \theta_t + \delta_{kt}.$$

Em que, IS é a importância segurada média. Tal medida é um *offset*, pois não lhe é atribuída nenhum coeficiente. Os efeitos fixos, incluindo o intercepto, são os fatores sexo do condutor e faixa etária do condutor e são modelados da forma $\beta_j \sim \mathcal{N}(0, 100^2)$. Temos ainda que u_k e v_k representam o efeito espacial estruturado (espacialmente correlacionado) e não estruturado (não correlacionado) dos estados, respectivamente. O termo u_k é definido condicionalmente como

$$u_k|\boldsymbol{u_{-k}}, \tau \sim \mathcal{N}\left(\frac{1}{n_k}\sum_{l\in\partial_k} u_l; \frac{1}{n_k\tau_u}\right),$$

em que ∂_k é o conjuntos de estados vizinhos do estado k que são os n_k estados que compartilham uma fronteira em comum com o estado k. Tal especificação, é denominada como modelo autorregressivo condicional intrínseco (iCAR) (BESAG; KOOPERBERG, 1995). Para o efeito espacial não estruturado assume-se que $v_k \sim \mathcal{N}(0, 1/\tau_v)$. Da união do Modelo de Besag, u_k e o modelo espacial não estruturado v_k origina-se o modelo de Besag-York-Mollìè (BYM) (BESAG et al.; 1991).

Já γ_t e θ_t são as duas componentes de efeito temporal, estruturada e não estruturada, respectivamente. Para o efeito não estruturado, temos $\theta_t \sim \mathcal{N}(0, 1/\tau_\theta)$. Já para o efeito γ_t, foi modelado dinamicamente, através de um passeio aleatório ora de ordem 1 (RW1), ora de ordem 2 (RW2), definidos como

RW1

$$\gamma_t|\gamma_{t-1} \sim \mathcal{N}(\gamma_{t-1}, 1/\tau_\gamma)$$

.

RW2

$$\gamma_t|\gamma_{t-1}, \gamma_{t-2} \sim \mathcal{N}(2\gamma_{t-1} + \gamma_{t-2}, 1/\tau_\gamma)$$

. Há ainda o parâmetro δ_{kt} que é não estruturalmente modelado e representa a interação entre o espaço e tempo, assim, $\delta_{kt} \sim \mathcal{N}(0, 1/\tau_\delta)$.

Knorr-Held (2000) complementa que no modelo de Bernardinelli et al. (1995) várias modificações e extensões são possíveis, no entanto, recomenda omitir apenas efeitos principais em que sua interação não é assumida. Dessa forma, a Tabela 2 elenca algumas variações do modelo geral (1) a serem ajustados.

Tabela 1: Componentes aleatórios presentes no modelo

Modelo	u_k	v_k	γ_t	θ_t	δ_{kt}
I	–	–	–	–	–
II	–	sim	–	sim	sim
III	sim	–	–	sim	–
IV	sim	–	sim (RW1)	–	–
V	sim	–	sim (RW1)	sim	–
VI	sim	sim	–	sim	sim
VII	sim	sim	sim (RW1)	sim	sim
VIII	sim	sim	sim (RW2)	sim	sim

O software R (R CORE TEAM, 2019) foi utilizado para a análise estatística, enquanto que o mapa do Brasil, bem como, sua estrutura de vizinhança foram obtidos junto ao Ministério Meio Ambiente. O método INLA (do inglês, *Integrated Nested Laplace Approximation*) foi utilizado para obtenção das marginais a posteriori.

3 Resultados e Discussão

Avaliar a qualidade de um modelo é verificando a acurácia de suas predições. Uma maneira de fazer essa avaliação é comparando diferentes modelos e selecioná-lo baseado em algum critério de seleção. Com a finalidade de selecionar o modelo mais acurado, calculou-se o Critério de Informação de Watanabe-Akaike (WAIC), que está apresentado na Tabela 3.

Tabela 2: Seleção de Modelos via Critério de Informação Watanabe - Akaike

Modelo	WAIC	p_W
I	27633,91	8,16
II	22687,91	200,29
III	23702,71	47,50
IV	23702,72	47,49
V	23702.53	47,39
VI	**22687,81**	200,15
VII	22688,00	200,21
VIII	22687,97	200,19

Pela Tabela 2 observa-se que os modelos de maior destaque são os modelos II, VI e VIII, sendo o modelo VI o mais acurado dentre os modelos ajustados, uma vez que detém o menor $WAIC = 22687{,}81$. Adicionalmente, o modelo VI possui um menor número efetivo de parâmetros $p_W = 200{,}15$ contra 200,19 e 200,29 dos modelos VIII e II, respectivamente. Dessa forma, o modelo VI é mais acurado e mais parcimonioso que os seus demais concorrentes e, portanto, é o modelo selecionado para o estudo. Um resumo dos efeitos do modelo VI é apresentado na Tabela 3.

A Figura 1 apresenta a média a posteriori do efeito espacial para cada estado brasileiro. Destaques para os estados Rio de Janeiro e Pará (áreas mais escuras) que obtiveram

Tabela 3: Média e desvio padrão a posteriori e intervalo de credibilidade de 95% dos efeitos do modelo VI

$\boldsymbol{x}$	Média	sd	$q_{0.025}$	$q_{0.975}$	$\boldsymbol{\Theta}$	Média	sd	$q_{0.025}$	$q_{0.975}$
β_0^*	-3,1944	0,0260	-3,2464	-3,1424	ϕ	712,54	22,97	668,15	758,59
β_{masc}	0,1137	0,0016	0,1105	0,1169	τ_v	1826,48	1540,76	273,65	5967,1
β_{26a35}	-0,2273	0,0026	-0,2323	-0,2223	τ_u	94,53	32,6	47,34	174,12
β_{36a45}	-0,3108	0,0026	-0,3159	-0,3058	τ_θ	233,28	110,86	75,77	501,43
β_{46a55}	-0,3091	0,0026	-0,3141	-0,3041	τ_δ	853,15	100	671,82	1065,14
β_{55+}	-0,3594	0,0026	-0,3645	-0,3544					

$*$ Intercepto confundido com sexo feminino e idade de 18 a 25 anos

os maiores efeitos dentre os estados, o que implica em uma maior influência desses estados no aumento do valor médio do prêmio de seguro de automóveis. Por outro lado, Santa Catarina, Sergipe e Paraíba (áreas em branco) foram os estados que tiveram efeitos espaciais menores do que 1, de modo que contribuíram para uma diminuição do valor médio do prêmio em seus respectivos estados. Ademais, todos os estados da região sul e a maioria da região nordeste, tiveram efeitos espaciais menores que 1 e de uma maneira geral, tem-se que as regiões norte, centro-oeste e sudeste apresentaram efeitos maiores que 1, exceto Minas Gerais, Distrito Federal e Mato Grosso do Sul em que tais efeitos provocam uma redução no valor do prêmio médio.

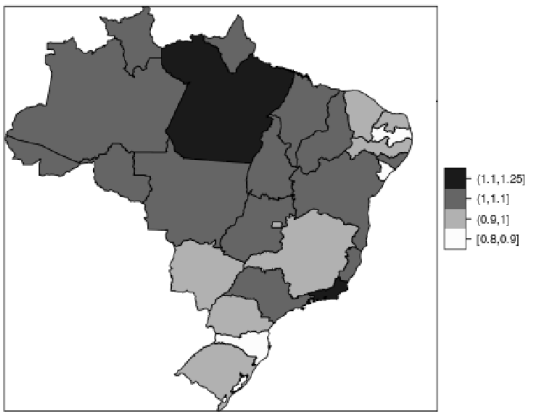

Figura 1: Média a posteriori para os efeitos espaciais para cada estado do Brasil ($\boldsymbol{\zeta} = exp(\boldsymbol{u} + \boldsymbol{v})$) referente aos anos de 2007 a 2014

A Figura 2 mostra a média a posteriori da tendência temporal para os anos de 2007 a 2014. De uma maneira geral, houve um efeito crescente no valor médio do prêmio entre os anos de 2007 a 2014, com destaque para as quedas 2008, 2009 e 2011 e para as ascensões em 2013 e 2014.

A Figura 3 mostra a média a posteriori da interação espaço-tempo para o valor do prêmio médio para os anos entre 2007 a 2014. Os estados com coloração mais escura em suas respectivas áreas, tiveram, para o ano específico, uma tendência temporal mais acentuada que a média da tendência temporal para o ano em questão. No ano de 2014, em que a média a posteriori da tendência temporal foi a mais alta dentre os anos considerados (ver Fig. 10) os estados como Rio de Janeiro, Espirito Santo, Amazonas, Bahia e todos os outros estados com coloração mais escura, tiveram uma média a posteriori da tendência temporal ainda mais acentuada do que a média geral da tendência temporal para o ano

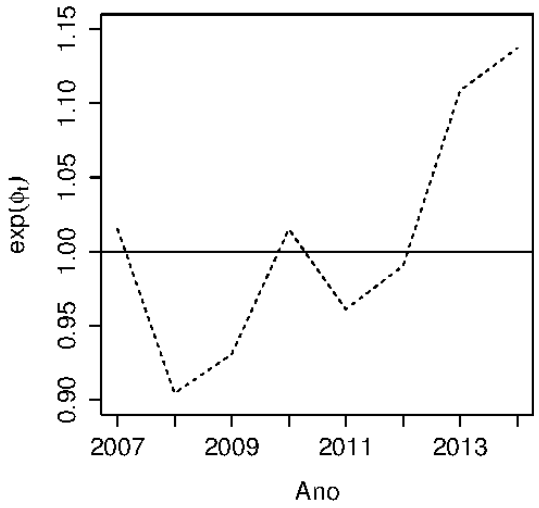

Figura 2: Tendência temporal da média a posteriori do valor médio do prêmio de seguro de automóveis no Brasil para os anos de 2007 a 2014

de 2014. Por outro lado, em 2014, estados como o de Santa Catarina, Rio Grande do Sul, Maranhão e outros que possuem a área com coloração mais clara, tiveram uma média a posteriori da tendência temporal menos acentuada que a média a posteriori da tendência temporal para o ano de 2014.

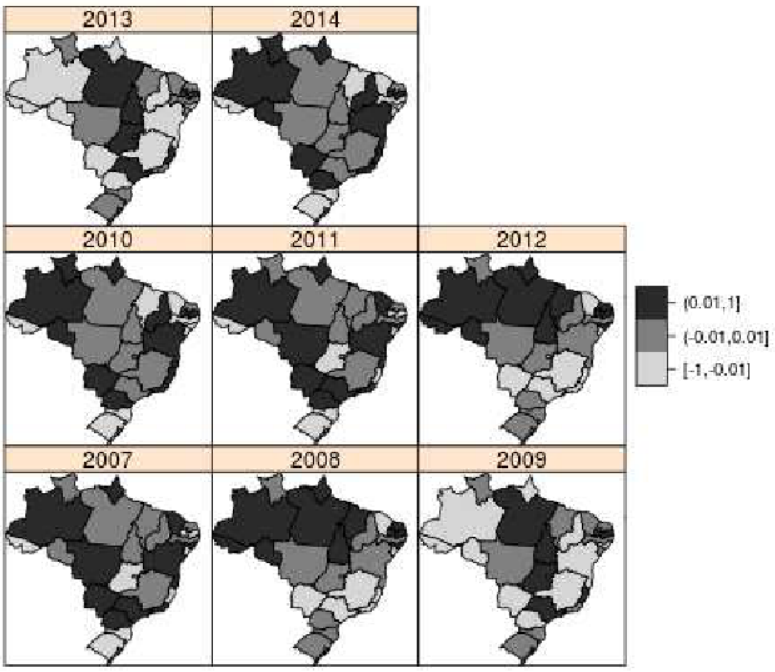

Figura 3: Média a posteriori da interação espaço-tempoδ_{kt} para o valor médio do prêmio de seguro de automóveis no Brasil para os anos de 2007 a 2014

Quanto ao modelo gama ajustado para o valor médio do prêmio de seguro de automóveis, cabe alguns destaques:

- O computador utilizado para os ajustes do modelo foi um com processador Intel Core I7 CPU 5500U @ 2,4 GHz $\times$ 4 e memória de 8 Gb DDR3 L. O tempo total

de execução para o modelo selecionado foi de aproxmadamente 8s utilizando a estratégia da aproximação gaussiana, 11s para a aproximação de Laplace simplificada e de 155s para a aproximação de Laplace (completa). Ademais, em termos inferenciais, não houve diferença entre as estimativas dos efeitos para o modelo ajustado com Aproximação de Laplace completa (mais precisa) e Aproximação Gaussiana (menos precisa).

- Verifica-se que condutores do sexo feminino tem uma redução no prêmio em relação aos condutores do sexo masculino. A classe de condutores do sexo masculino entre 18 e 25 anos sofrem um aumento de 12,04% em relação ao prêmio médio pago por condutores do sexo feminino na mesma faixa etária que, inicialmente, pagam um prêmio médio equivalente a 4,10% da importância segurada média. Condutores com mais idade possuem uma redução do valor de seu prêmio em relação às outras classes de condutores, sendo ainda maior comparado com os que possuem entre 18 e 25 anos.

4 Conclusões

Com os resultados apresentados, pode-se observar que a metodologia INLA fornece uma nova, rápida e interessante maneira de aproximar as marginais a posteriori para uma ampla variedade de modelos bayesianos hierárquicos e pode ser utilizada por vários pesquisadores de diferentes áreas de aplicação. Dessa forma, a metodologia INLA representa um importante avanço no campo da estatística bayesiana e, por ser mais veloz, surge como uma boa alternativa às aproximações via MCMC, para o caso em que o modelo ajustado pertença à classe de modelos latentes gaussianos.

Por fim, observou-se que a modelagem que inclui efeitos agrupados (aleatórios) foi capaz de descrever melhor os dados de seguro de automóveis do que o modelo linear generalizado. E dentre os modelos de efeitos agrupados, os mais precisos incluem efeitos espaciais.

Agradecimentos

Agradecemos à CAPES e à FAPEMIG pelo apoio financeiro.

Referencias Bibliográficas

ANTONIO, K.; BEIRLANT, J. Actuarial statistics with generalized linear mixed models, *Insurance: Mathematics and Economics*. 40(1): 58? 76, 2007.

BERNANRDINELLI, L., et al. Bayesian analysis of space-time variation in disease risk, *Statistics in medicine* 14(21-22): 2433? 2443., 1995.

DICKSON, D. C. *Insurance risk and ruin*, 1 edn, Cambridge University Press, United Kingdom, 2010

FAHRMEIR, L., TUTZ, G. *Multivariate statistical modelling based on generalized linear models*, 1 ed., Springer Science & Business Media, 2013.

McCULLOCH, C. E., SEARLE, S. R. and NEUHAUS, J. M. *Generalized, Linear, and Mixed Models,* v. 2, 2 edn, Wiley, New York, NY, 2008.

R CORE TEAM. *R*: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2012. ISBN 3-900051-07-0, URL `http://www.R-project.org/`.

RUE, H.; HELD, L. *Gaussian Markov random fields: theory and applications*, 1 ed, CRC press.