Universidade Federal do Rio de Janeiro

Instituto de Matemática

Bacharelado em Ciências Atuariais e Estatística

Modelos Lineares Generalizados Aplicado a Modelagem da Frequência de Sinistros

Jéssica Vitória dos Santos Areias

Rio de Janeiro

2023

Modelos Lineares Generalizados Aplicado a Modelagem da Frequência de Sinistros

Jéssica Vitória dos Santos Areias

Universidade Federal do Rio de Janeiro
Instituto de Matemática
Bacharelado em Ciências Atuariais e Estatística

Orientadora: $Prof^a.Dr^a$ Viviana das Graças Ribeiro Lobo

Rio de Janeiro

2023

Modelos Lineares Generalizados Aplicado a Modelagem da Frequência de Sinistros

Jéssica Vitória dos Santos Areias

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE MATEMÁTICA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIENCIAS ATUARIAIS E EM ESTATÍSTICA.

Aprovada por:

$Prof^a.Dr^a$ Viviana das Graças Ribeiro Lobo
IM-UFRJ

Prof. Dr João Batista de Morais Pereira
IM-UFRJ

Prof. Dr William Lima Leão
IM-UFRJ

Rio de Janeiro, RJ - Brasil

2023

CIP - Catalogação na Publicação

A679m

Areias, Jéssica Vitória dos Santos
Modelos lineares generalizados aplicado a modelagem da frequência de sinistros / Jéssica Vitória dos Santos Areias. -- Rio de Janeiro, 2023.
48 f.

Orientadora: Viviana das Graças Ribeiro Lobo.
Trabalho de conclusão de curso (graduação) - Universidade Federal do Rio de Janeiro, Instituto de Matemática, Bacharel em Ciências atuariais, 2023.

1. Frequência de sinistros. 2. Modelos lineares generalizados. 3. Inferência Bayesiana. I. Lobo, Viviana das Graças Ribeiro, orient. II. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

Agradecimentos

Em primeiro lugar, agradeço a Deus, pois Ele me permitiu, mesmo com todas as adversidades encontradas no caminho, chegar até aqui. Além de permitir, Ele me estendeu a mão e ajudou em diversas situações na qual eu achava que não seria mais possível seguir em frente.

No início desse sonho, que era ingressar na famosa UFRJ, havia uma pessoa que lutava tanto quanto eu para que ele se tornasse realidade, minha mãe Bárbara. Não tem como chegar até aqui e não pensar nela, que vibrou até mais do que eu mesma quando viu meu nome na lista de aprovados. Se tem alguém que batalhou para que tudo isso fosse possível, esse alguém foi você, mãezinha, e eu não tenho palavras para te agradecer por tudo. Sei que você estaria vibrando tanto quanto eu e espero que do céu, você consiga sentir essa felicidade e receber os meus mais sinceros agradecimentos e todo o amor que vai estar para sempre guardado em meu coração.

Agradeço ao meu pai João, que diante de todas as dificuldades, sempre se demonstrou forte e me amparou em inúmeros momentos de fragilidade. Obrigada pai, por todo trabalho árduo para investir em mim, por ser tão incansável e por acreditar que eu chegaria até aqui. Tudo isso só foi possível, porque você sempre esteve presente para nós.

Não são todas as pessoas que têm a sorte de ter uma melhor amiga que é desde sempre e para sempre. Na vida, fazemos amizades que vêm e vão, mas existe aquela amiga que sempre fica, essa é a minha irmã Gabrielle. Quando você nasceu, eu sabia que teria uma irmã/amiga para a vida inteira. Obrigada por todo cuidado comigo, pelos lanches nas minhas madrugadas de estudo, por me aturar nos meus piores dias, pelas palavras que sempre me reergueram, obrigada por sempre estar comigo e por trazer a nossa família um motivo a mais para continuar seguindo, a minha linda sobrinha Isabela; tão pequena e já faz tanto por nós, obrigada minha neném.

Agradeço ao meu namorado André, Deus pensou em mim com tanto carinho, que me deu de presente um homem maravilhoso. Obrigada pelo cuidado, pelo apoio, por todas as vezes que me acompanhou até a faculdade; obrigada por sempre acreditar em mim, me incentivar e por todos os abraços que acalmaram meu coração e me ajudaram a

continuar seguindo em frente. E claro, não poderia esquecer das tortinhas que alegravam até os meus dias mais estressantes, muito obrigada.

Na vida, quem tem família, tem tudo, por isso eu não poderia deixar de citar minha madrinha Laura, minha tia Jocilene e meu primo Anselmo. Obrigada dinda, por estar sempre presente, por todas as deliciosas marmitas que fazia para mim e por todo carinho e amor. Obrigada tia, por todos os conselhos e por me ouvir todas as vezes que eu chegava na sua casa e precisava desabafar sobre tudo, seja relacionado a faculdade ou não; você é para mim um exemplo de força que eu sempre seguirei. Obrigada meu primo, você é como meu irmão mais velho, obrigada pelos conselhos, por toda ajuda e por todos os ensinamentos que me fazem querer, um dia, ser tão inteligente quanto você.

Agradeço aos meus sogros Margareth e Luis Augusto, que desde sempre me acolheram como uma filha, me incentivaram e sempre acreditaram em mim. Minha sogra, obrigada por todas as orações, tenho a certeza que se cheguei aonde estou, é porque intercedeu por mim. Meu sogro, obrigada pelas inúmeras vezes que, às 6h da manhã, levantou de sua cama para nos levar a faculdade, prezando pelo nossa segurança.

Agradeço a todos os familiares que, de alguma forma, me ajudaram a chegar até aqui. Saibam que todas as atitudes, citadas aqui neste agradecimento e também as não citadas, foram de grande importância para mim.

A UFRJ me proporcionou muitas experiências boas, e eu não poderia viver elas sozinha. Obrigada Geysa e Gabriela, por me acolherem tão bem no primeiro momento de início de graduação, que é o mais difícil de todos; obrigada por todas as trocas e todo carinho. Obrigada Mariana Soares pelas diversas ajudas não só nos trabalhos de todos os cursos que fizemos juntas, mas também por ter me ajudado tanto e também me acolhido nesse trabalho final de conclusão de curso. Obrigada Larissa, por ser minha amiga de todos os momentos, por sempre me ajudar a prosseguir e por todas as trocas nos estudos também. Obrigada Mariana Rumma, por ser sempre tão doce e ter sempre uma palavra de conforto.

Por fim, muito obrigada a todos os professores que, com todos os seus ensinamentos, me ajudaram a chegar até aqui. Obrigada João e William, pelo prazer que me proporcionaram ao aceitarem meu convite. E um obrigada especial a Viviana, que aceitou ser minha orientadora e me auxiliou nessa reta final.

Resumo

No ramo dos seguros de automóveis, as companhias de seguros necessitam identificar o risco de cada um dos seus clientes sofrerem um acidente, de forma a calcular o prêmio adequado que o cliente deve pagar. Se o valor do prêmio for inferior aos custos que a companhia terá como obrigação em caso de sinistro, esta incorre em perdas financeiras. Surge assim a necessidade de estudar as características que influenciam os acidentes.

Diante disso, este trabalho tem o objetivo de estudar a ocorrência de sinistros em função de variáveis relacionadas ao segurado, como sexo, faixa etária e categoria do carro, de forma independente para cada estado do Brasil.

Como a variável resposta corresponde uma contagem, laçamos mão de Modelos Lineares Generalizados, que permitem a modelagem de dados não necessariamente normais. Especificamente, consideremos a distribuição de Poisson e o procedimento de inferência foi feito sob o enfoque bayesiano.

Palavras-Chave: Frequência de Sinistros; Modelos Lineares Generalizados; Inferência Bayesiana.

Abstract

In the field of car insurance, insurance companies need to assess the risk of each of their clients being involved in an accident in order to calculate the appropriate premium the client should pay. If the premium amount is lower than the costs the company would be obliged to cover in case of an accident, it incurs financial losses. This leads to the necessity of studying the factors that influence accidents.

Therefore, this work aims to study the occurrence of accidents based on variables related to the insured party, such as gender, age group, and car category, independently for each state in Brazil.

As the response variable corresponds to a count, we employ generalized linear models, which allow for modeling of non-normally distributed data. Specifically, let's consider the Poisson distribution, and the inference procedure was carried out under the Bayesian approach.

Keywords: Frequency of Claims, Generalized Linear Models, Bayesian inference.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O seguro mais utilizado no Brasil é o de automóveis, mesmo assim, apenas 30% da frota do país é segurada (cerca de 20 milhões de veículos), de acordo com a Confederação Nacional das Empresas de Seguros Gerais, Previdência Privada e Vida, Saúde Suplementar e Capitalização (CNseg)CNseg (2020). Isso demonstra que a população tem uma preocupação adicional com o carro, que por estar na rua na maior parte do tempo, nos faz pensar que está mais exposto ao risco.

Os sinistros podem ser definidos como a ocorrência de todo e qualquer evento que possua uma cobertura em um dado seguro contratado e esteja devidamente especificado na apólice. Existem diversos tipos de seguros para automóveis, como por exemplo seguro contra colisão parcial, colisão total, furto/roubo, incêndio ou até mesmo seguros classificados dentro de uma categoria que cobre diversos tipos de sinistros diferentes, assim como a cobertura de assistência 24h.

O seguro de automóveis permite ao segurado se resguardar, caso ocorra algum desfortúnio com o seu veículo, um bem que é considerado valioso e que possui alto custo financeiro. Dessa forma, ele possui o direito de receber uma indenização, caso ocorra sinistro. A proteção contra algum tipo de incidente só é possível mediante ao pagamento de um prêmio (importância paga pelo segurado a seguradora em troca da transferência do risco a que ele está exposto).

Nesse sentido, é fundamental para uma seguradora detectar quais variáveis influenciam na frequência de sinistros durante a vigência de um contrato, variável importante para o cálculo do prêmio que o segurado deverá pagar, já que ela precisa gerar um fundo com os valores pagos pelos contratantes e assim conseguir cumprir com as suas obrigações e se manter solvente no mercado.

Dito isso, este trabalho visa modelar a frequência de sinistros utilizando Modelos Lineares Genealizados, a metodologia introduzida por Nelder e Wedderburn (1972).

Os Modelos Lineares Generalizados (GLM's) consistem em uma classe de modelos de regressão mais abrangente que o Modelo Linear Normal Clássico, que permitem analisar dados não normais (ou não gaussianos), podendo ser utilizada em diversas áreas de estudo, como no caso deste trabalho. Esses modelos podem ser utilizados no cálculo da frequência de sinistros, por exemplo.

O objetivo geral dessa pesquisa é utilizar a abordagem dos Modelos Lineares Generalizados, testando possíveis distribuições como a Poisson, já que a variável número de sinistros se trata de uma contagem, e assim estudar fatores que podem interferir na ocorrência de sinistros em uma determinada região, como sexo e faixa etária, e também verificar o comportamento das regiões em relação a frequência de sinistros.

Uma tese que utilizou os Modelos Lineares Generalizados para modelagem da frequência de Sinistros foi o Ferreira (2013), utilizando dados de uma seguradora específica, na qual foram modelados por uma distribuição Poisson, obtendo bons resultados. Desta forma, utilizamos esta tese como base para construção deste trabalho final.

O conteúdo deste trabalho está organizado em 5 capítulos. Após a contextualização, apresentada neste capítulo, é mostrada a análise exploratória dos dados no capítulo 2. A teoria dos Modelos Lineares Generalizados utilizada neste trabalho é dissertada no capítulo 3, e por fim é apresentada a análise dos resultados e as considerações finais, juntamente com as sugestões para trabalhos futuros nos capítulos 4 e 5.

Capítulo 2

Análise Exploratória de Dados

2.1 Introdução à Base de Dados

O estudo de uma base de dados confiável para uma melhor visualização de fatores, como: sexo, idade, renda, entre outras características do segurado, que podem ou não influenciar na ocorrência de sinistros, é de extrema importância para uma seguradora, já que ela necessita ter estimativas (probabilidade de morte) bem definidas para a realização do cálculo do prêmio de forma a não prejudicar o lucro da empresa; e deste modo obter sucesso em seu objetivo.

Pensando nesta perspectiva, os dados utilizados foram retirados do site da Autoseg - Sistema de Estatísticas de Automóveis da SUSEP, conforme AUTOSEG (2020), que permite acessar o banco de dados completo ou até mesmo efetuar consultas on-line.

As informações apresentadas no site provêm de arquivos enviados semestralmente pelas companhias seguradoras que atendem ao item 9 do Manual de Orientação anexo à Circular SUSEP nº 522/2015, os quais incluem dados referentes a apólices vigentes e sinistros ocorridos no período de análise. O manual de orientação encontra-se disponível no site oficial da SUSEP, conforme SUSEP (2020).

O sistema era atualizado semestralmente, até 2020, última atualização. O banco de dados utilizado foi o do segundo semestre de 2020, uma vez que é o mais completo atualmente, contendo dados desde o segundo semestre de 2006 até o segundo semestre de 2020.

Na subseção abaixo segue a descrição mais detalhada das variáveis utilizadas.

2.1.1 Variáveis de Interesse

O sistema fornece informações sobre número de veículos expostos, prêmio médio e número de sinistros, classificadas de acordo com a categoria do veículo, região ou CEP de circulação, e perfil do segurado.

1. **Exposição:** O conceito de exposição leva em conta o tempo em que cada apólice esteve vigente, dentro do período semestral observado em cada atualização do Autoseg.

2. **Prêmio Médio:** Da mesma forma que a IS Média, o prêmio médio representa a média dos prêmios das apólices incluídas no grupamento, ponderada pela exposição de cada uma delas.

3. **Frequência de Sinistros:** A frequência é a quantidade de sinistros de incêndio, roubo, colisão e outras causas, por apólice.

4. **Categoria do Veículo:** A variável categoria dispõe de duas modalidades, diferenciando os automóveis entre passeio nacional e passeio internacional.

5. **Região:** A região é definida pelas 27 unidades federativas que compõem o Brasil, sendo 26 estados e o Distrito Federal.

6. **Sexo do Segurado:** Identificação do gênero do segurado, estruturado entre masculino e feminino.

7. **Faixa Etária do Segurado:** Identificação da faixa etária do segurado, composta po 5 tipos de faixa diferentes: entre 18 e 25 anos, entre 26 e 35 anos, entre 36 e 45 anos, entre 46 e 55 anos e acima de 55 anos.

Nos dados disponibilizados pela Autoseg, a frequência de sinistros é dividida por tipos de cobertura, são elas:

- **FREQ_SIN1:** Quantidade de sinistros das coberturas roubo ou furto;
- **FREQ_SIN2:** Quantidade de sinistros da cobertura colisão parcial;
- **FREQ_SIN3:** Quantidade de sinistros da cobertura colisão perda total;
- **FREQ_SIN4:** Quantidade de sinistros da cobertura incêndio;

- **FREQ_SIN9:** Quantidade de sinistros de outras coberturas, como assistência 24 horas, entre outras.

O tipo de cobertura definida como "outras coberturas"engloba os sinistros não incluídos nos dados de incêndio, roubo, ou colisão, ou seja, assistência 24 horas e outras coberturas como vidros, blindagem, equipamentos acessórios, etc.

2.2 Análise para Cada Cobertura

Em um estudo inicial, foi realizada uma análise da frequência relativa de sinistros por estado, para cada tipo de cobertura.

Para se obter um resultado mais fidedigno, é necessário considerar a quantidade de segurados expostos em cada região, por isso foi utilizada a frequência relativa, que é definida como:

$$Frequência\ Relativa = \frac{Frequência\ Absoluta\ de\ Sinistros\ em\ cada\ região}{Quantidade\ de\ Expostos\ em\ cada\ região}$$

A Figura 2.1, mostra os mapas referentes a frequência relativa de sinistros, em cada estado do Brasil, para todas as coberturas citadas anteriormente. As escalas dos mapas estão diferentes, e por isso foi adicionado uma legenda com a frequência relativa de sinistros para cada estado em cada categoria a fim de que seja possível realizar uma comparação justa e correta.

É possível notar que em todas as coberturas a frequência de sinistros é maior na região Sudeste do país, área esta que contém os estados mais populosos do Brasil, como: o município de São Paulo com 12,3 milhões de pessoas e o município do Rio de Janeiro com 6,7 milhões de pessoas; dados retirados do site oficial de notícias do IBGE para o ano de 2020, conforme IBGE (2020).

Adicionalmente, o tipo de cobertura "outras coberturas"é o que mais se destaca em relação as outras; quando o enfoque é o número de ocorrência de sinistros nos estados, pode-se ver que ela possui frequências muito mais altas em praticamente todas as áreas do país.

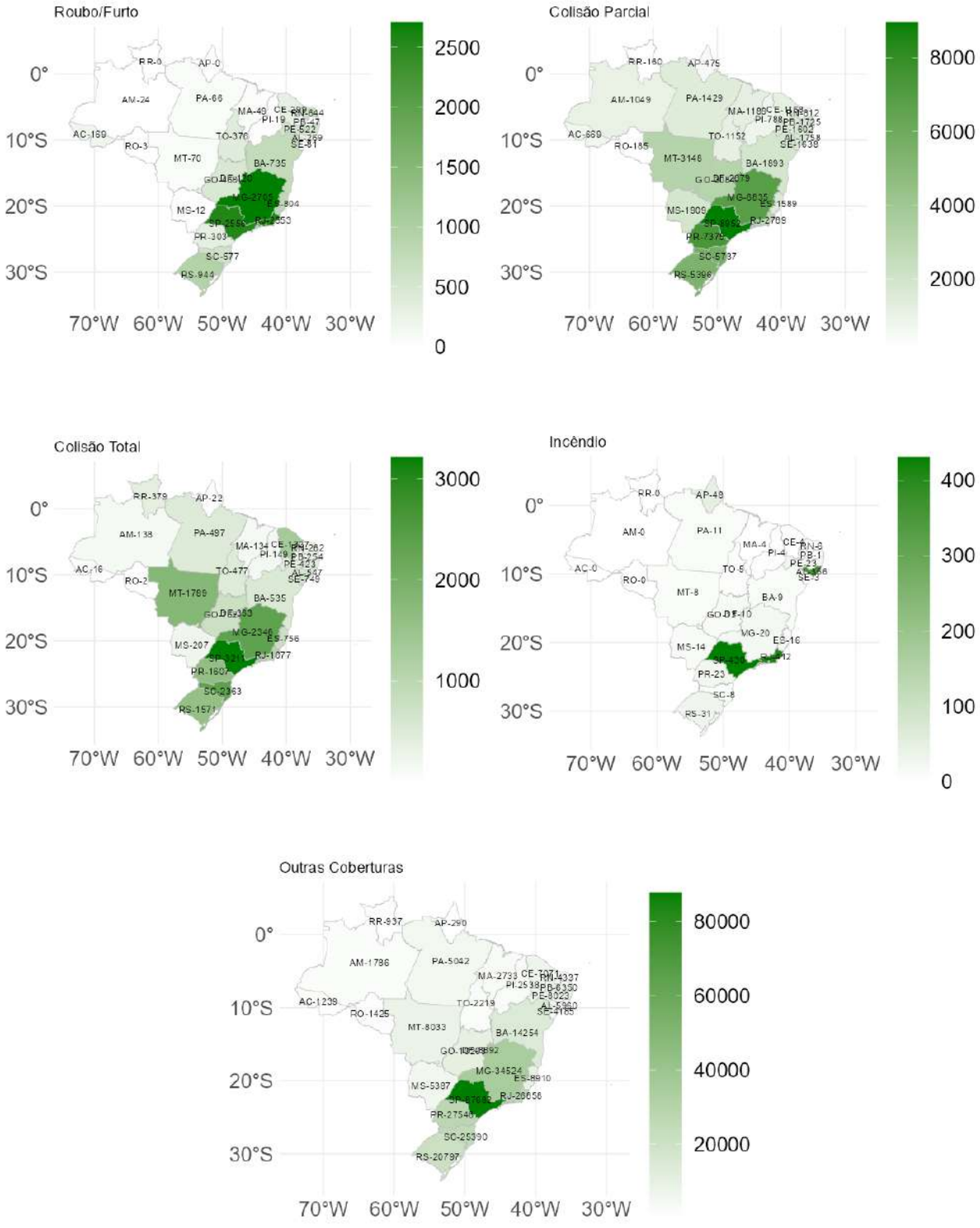

Figura 2.1: Mapas da Frequência Relativa de Sinistros para as Coberturas

Foi utilizada uma medida de porcentagem, na qual soma-se todas as frequências relativas de sinistros para cada cobertura e divide pelo total da frequência de sinistros ocorridos para todas as coberturas, os resultados obtidos constam na Tabela 2.1 abaixo apresentada, e desta forma é possível confirmar o que já foi visto nos mapas.

De fato, com 76% de frequência relativa de sinistros, o tipo "outras coberturas"se sobressai de forma significativa e por isso, foi a cobertura escolhida para a análise e aplicação dos métodos utilizados neste trabalho.

Cobertura	Porcentagem
Roubo e Furto	3,29%
Colisão Parcial	15,35%
Colisão Total	5,02%
Incêndio	0,33%
Outras Coberturas	76,00%
Total	100,00%

Tabela 2.1: Porcentagem de frequência relativa de sinistros para cada cobertura.

2.3 Análise para Outras Coberturas

De agora em diante, o trabalho será relativo somente ao tipo de cobertura "outras coberturas", por motivos já mencionados na seção anterior.

Com o propósito de uma análise mais profunda dos dados, foi feito um estudo da frequência de sinistros em cada região do país, separado por sexo, faixa etária dos segurados e categoria do carro. Para isso, a faixa etária foi dividida somente em duas: entre 18 e 35 anos e acima de 35 anos. Isso foi feito, pois após plotar diversos mapas para as diferentes faixas etárias existentes, nota-se que a diferença de frequência de sinistros para faixas etárias acima de 35 anos (entre 36 e 45 anos, entre 46 e 55 anos e acima de 55 anos) não se mostraram serem tão relevantes quanto a diferença de frequência de sinistros acerca das faixas etárias entre 18 e 35 anos e acima de 35 anos, por isso agregamos as últimas idades.

Observa-se, na Figura 2.2, que tanto para o sexo feminino como para o masculino, segurados(as) com idades acima de 35 anos cometem mais sinistros do que segurados(as) com idade entre 18 e 35 anos. Já quando compara-se os sexos, mulheres parecem cometer mais sinistros do que os homens, mesmo apresentando uma diferença pouco visível.

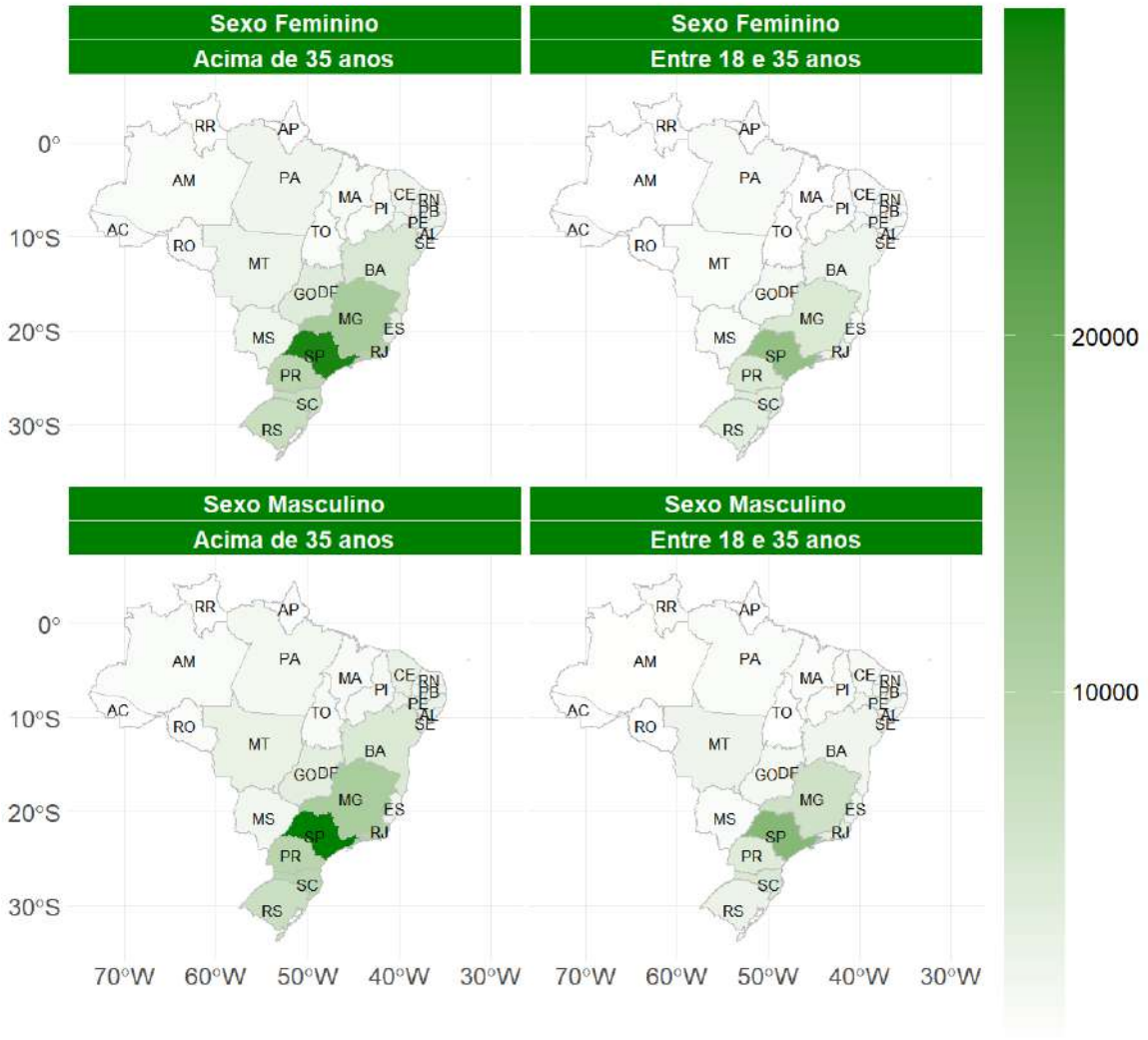

Figura 2.2: Mapa da Frequência Relativa de Sinistros para Outras Coberturas por Sexo e Faixa Etária.

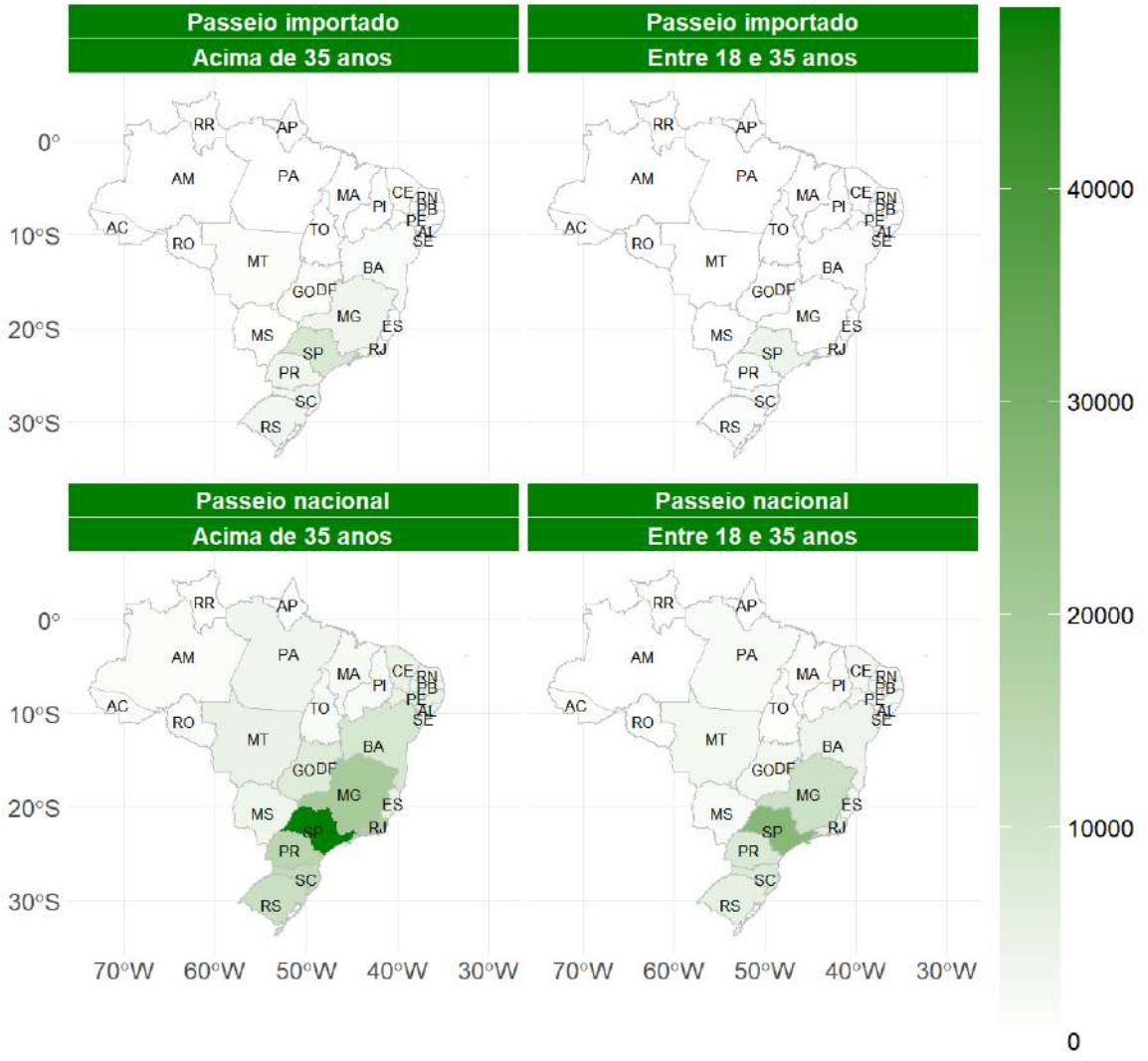

Figura 2.3: Mapa da Frequência Relativa de Sinistros para Outras Coberturas por Categoria do Veículo e Faixa Etária.

Na Figura 2.3, nota-se que há uma frequência de sinistros muito maior para segurados que possuem carro de passeio nacional. E também, é possível ver que, para ambos os tipos de veículos, segurados acima de 35 anos têm maior frequência de sinistros do que beneficiários com idades entre 18 e 35 anos.

Foi feito um gráfico utilizando as variáveis sexo e categoria do veículo, porém não observou-se uma diferença significativa, e por isso não foi apresentado.

2.3.1 Análise de Autocorrelação Espacial

Segundo Banerjee et al. (2014) pesquisadores em diversas áreas, como climatologia, ecologia, saúde ambiental e marketing imobiliário, enfrentam cada vez mais a tarefa de analisar dados que são: altamente multivariados, geograficamente referenciados e frequentemente apresentados como mapas, e correlacionados temporalmente, como em estruturas de séries temporais. Como, por exemplo, pode-se citar o estudo de Oliveira (2008) que teve como objetivo realizar uma análise espacial da criminalidade no estado do Rio Grande do Sul. No modelo, a criminalidade nas cidades pode ser explicada por características locais em que o ambiente, a vizinhança e o histórico do indivíduo afetam a criminalidade; são utilizados dados municipais agregados para homicídios, roubos e furtos no ano de 2000.

Ao considerar a informação da localização do dado disposto, utilizamos o que chamamos de dados espaciais, que são dados de uma variável associados a uma coordenada espacial. Estes dados podem possuir o que chamamos de autocorrelação espacial, que pode ser definida como a tendência de que o valor de uma variável, associada a uma determinada localização, assemelha-se mais aos valores de suas observações vizinhas do que ao restante das localizações do conjunto amostral. Como os nossos dados aparentam apresentar uma autocorrelação espacial, resolvemos fazer uma análise acerca disso.

Para introduzir a associação espacial, é denifido uma estrutura de vizinhança com base nos arranjos dos blocos (regiões) no mapa, e para analisar essa dependência, existem diversos indicadores de autocorrelação espacial, como: o índice de Moran e o índice de Geary, segundo Banerjee et al. (2014). Tais indicadores podem ser globais ou locais e serão utilizados neste trabalho.

Matriz de Proximidade

Uma ferramenta muito útil na exploração de dados de unidade de área e essencial para a utilização dos índices apresentados acima é a matriz de proximidade, que denotaremos

como $\boldsymbol{W}$. Segundo Banerjee et al. (2014), nos modelos de dados de área, as regiões geográficas ou quarteirões (CEP, municípios, etc.) podem ser denotados por Y_i, onde n é o número de unidades de área apresentadas na base de dados, ou seja, o número de regiões; desta forma, as entradas w_{ij} da matriz $\boldsymbol{W}$ conectam espacialmente as regiões i e j, de forma que w_{ii} é definido como 0, uma vez que uma região não pode ser vizinha dela mesma. As possibilidades de definição para os fatores "ser vizinho"ou "não ser vizinho"podem ser binárias, na qual $w_{ij} = 1$ se as regiões i e j possuem algum limite comum e $w_{ij} = 0$ caso contrário. De forma alternativa, w_{ij} poderia refletir a distância entre as unidades. Neste trabalho, será utilizada a matriz definida como binária, posto que, atualmente, é a mais encontrada na literatura.

Considere a Figura 2.4, onde um retângulo representa uma região composta por 6 microrregiões chamadas Y_1, Y_2, Y_3, Y_4, Y_5 e Y_6. Há dois critérios existentes para a definição de duas regiões serem ou não vizinhas, ambos baseados no jogo de dama; uma delas é chama de critério torre, na qual são consideradas vizinhas, regiões que dividem fronteira, por exemplo, Y_1 é vizinho de Y_2 e de Y_4, mas não é vizinho de Y_5. Já o outro critério é chamado de rainha, onde Y_1 é vizinho de Y_2, Y_4 e também de Y_5, já que a rainha possui permissão para transitar na diagonal. O critério mais utilizado é o chamado de torre e por isso vamos utilizá-lo neste trabalho.

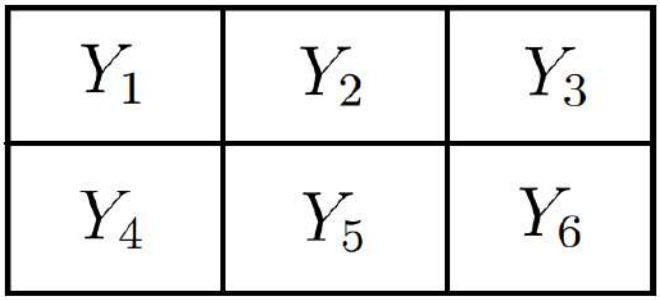

Figura 2.4: Mapa de Regiões Ilustrativo e Simplificado.

Abaixo, segue a matriz de proximidade, também conhecida como matriz de pesos, para ilustração:

	Y1	Y2	Y3	Y4	Y5	Y6
Y1	0	1	0	1	0	0
Y2	1	0	1	0	1	0
Y3	0	1	0	0	0	1
Y4	1	0	0	0	1	0
Y5	0	1	0	1	0	1
Y6	0	0	1	0	1	0

aqui $w_{ij} = 0$ indica que as regiões não são vizinhas e $w_{ij} = 1$ indica que as regiões são vizinhas, conforme já explicado anteriormente.

Medidas de Associação Espacial

Existem duas estatísticas muito utilizadas para medir a associação espacial entre unidades de área, chamadas de Índice I de Moran e o Índice C de Geary.

Índice I de Moran (Global)

O Índice I de Moran assume a seguinte forma, segundo Banerjee et al. (2014):

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(Y_i - \overline{Y})(Y_j - \overline{Y})}{\left(\sum_{i\neq j} w_{ij}\right)\sum_{i=1}^{n}(Y_i - \overline{Y})^2},$$

em que: Y_i é o valor do atributo da i-ésima área observada; Y_j é o valor do atributo da j-ésima área observada; $\overline{Y}$ é o valor médio do atributo da região de estudo e w_{ij} é a matriz de proximidade espacial normalizada.

O valor do índice de I de Moran deve estar no intervalo de [-1,1]. Quanto mais próximo de 1, maior a similaridade da variável em questão entre as áreas próximas; quanto mais próximo de 0 menor a similaridade; e se I for menor do que 0, indica que a área possui dados inversamente correlacionados.

Este índice é uma medida global da autocorrelação espacial, pois indica o grau de associação espacial presente no conjunto de dados como um todo.

Essa estatística nos permite observar os resultados utilizando gráficos quando lidamos com o I de Moran local. Um deles é chamado de diagrama de espalhamento de Moran, que possibilita visualizar a dependência espacial; há também o mapa de *cluster* e o mapa de significância. Essas ferramentas serão explicadas posteriormente.

Índice I de Moran (Local)

A estatística de I de Moran global fornece uma análise de um agrupamento de observações como um todo, ou seja, é uma característica de um padrão espacial completo e não fornece uma indicação da localização desses agrupamentos.

Foi sugerido por Anselin (1995) um indicador local de associação espacial chamado de LISA, que possui como objetivo identificar aglomerados locais e outliers espaciais locais. Tal princípio possui duas características consideradas importantes. Primeiro, fornece uma estatística para cada local, com uma avaliação de significância. E segundo, estabelece uma relação proporcional entre a soma das estatísticas locais e uma estatística global correspondente.

Como já visto anteriormente, a estatística I de Moran global é expressa como uma soma dupla sobre os índices i e j, na qual w_{ij} é a matriz de pesos. A forma local desta estatística seria a soma da expressão em relação ao índice j para cada observação (localização) i fixada.

$$I = \frac{(Y_i - \overline{Y}) \sum_{j=1}^{n} w_{ij}(Y_j - \overline{Y})}{\left(\sum_{i \neq j} w_{ij}\right) \sum_{i=1}^{n} (Y_i - \overline{Y})^2},$$

onde $\frac{(Y_i - \overline{Y})}{\left(\sum_{i \neq j} w_{ij}\right) \sum_{i=1}^{n} (Y_i - \overline{Y})^2}$ é considerada fixa, pois é referente ao índice i.

A significância do I de Moran pode ser baseada em uma aproximação analítica, mas segundo Anselin (1995) não é muito confiável. Consultar Anselin (1995) para mais informações.

Uma melhor abordagem consiste em um método de permutação condicional, na qual uma observação (localização) é fixada e as demais observações são permutadas aleatoriamente para produzir uma distribuição de referência para a estatística local (uma para cada local).

O I de Moran local funciona da mesma forma que o global, exceto que a permutação é realizada para cada observação por vez, resultando em um p-valor para cada local, que pode ser utilizado para avaliar a significância.

O diagrama de espalhamento de Moran, é uma maneira adicional de visualizar a dependência espacial. Este gráfico, mostrado na Figura 2.5, é construído com base nos valores normalizados, onde o eixo da variável X representa os atributos e o eixo da variável

W_X representa a média dos atributos dos vizinhos. Com ele, é possível comparar os valores normalizados do atributo numa área com a média dos seus vizinhos.

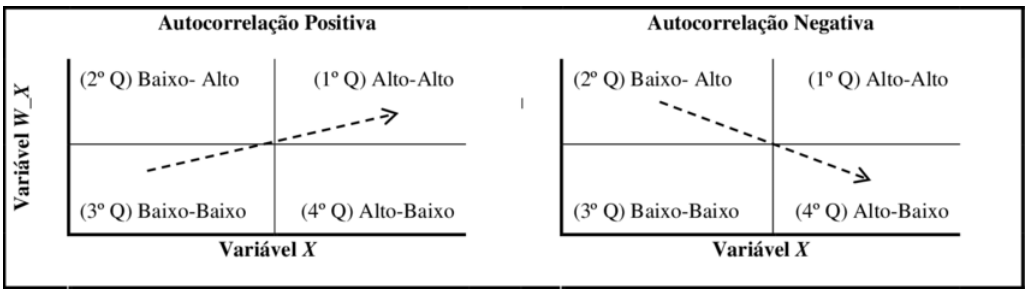

Figura 2.5: Esquema do Diagrama de Dispersão de Moran

O gráfico é dividido em 4 quadrantes:

- 1°Q (valores positivos e médias positivas) e 3°Q (valores negativos e médias negativas): indicam pontos de associação espacial positiva, no sentido de que uma localização possui vizinhos com valores semelhantes.
- 2°Q (valores negativos e médias positivas) e 4°Q (valores positivos e média negativas): indicam pontos de associação espacial negativa, no sentido de que uma localização possui vizinhos com valores distintos, indicando pontos de transições entre diferentes padrões espaciais ou pontos de não estacionariedade do atributo.

Adicionalmente, o gráfico de dispersão de Moran fornece uma classificação da associação espacial em quatro categorias, correspondendo à localização dos pontos nos quatro quadrantes do gráfico. Essas categorias são referidas como Alto-Alto (valores altos e médias altas), Baixo-Baixo (valores baixos e médias baixas), Baixo-Alto (valores baixos e médias altas) e Alto-Baixo (valores altos e médias baixas), em relação à média, que é o centro do gráfico. Pode-se construir ainda, um mapa de cluster usando as categorias citadas acima e um mapa de significância, que indicará quais regiões possuem autocorrelações que devem, de fato, serem consideradas. Ambos possuem uma conexão com o gráfico de dispersão.

Índice C de Geary

O Índice C de Geary assume a seguinte forma, conforme Banerjee et al. (2014):

$$C = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(Y_i - Y_j)^2}{2\left(\sum_{i\neq j} w_{ij}\right)\sum_{i=1}^{n}(Y_i - \overline{Y})^2},$$

em que: Y_i é o valor do atributo da i-ésima área observada; Y_j é o valor do atributo da j-ésima área observada; $\overline{Y}$ é o valor médio do atributo da região de estudo e w_{ij} é a matriz de proximidade espacial normalizada.

O índice C de geary nunca é negativo e tem média igual a 1. O valor de seu índice varia aproximadamente entre 0 e 2; valores de C no intervalo de 0 e 1 indicam associação espacial positiva, enquanto valores significativamente maiores que 1 ilustram aumento da autocorrelação espacial negativa.

O C de Geary é inversamente relacionado ao I de Moran global, mas não é idêntico. Embora o I de Moran e o C de Geary sejam medidas de autocorrelação espacial global, eles são ligeiramente diferentes. O C de Geary usa a soma das distâncias ao quadrado, enquanto o I de Moran usa covariância espacial padronizada. Ao usar distâncias quadradas, o C de Geary é menos sensível a associações lineares e pode captar autocorrelação onde o I de Moran não pode.

2.3.2 Aplicação das Medidas de Associação Espacial

Segundo PAIVA (2007), podemos utilizar teste de hipóteses para verificar a significância em uma análise espacial, utilizando o p-valor encontrado nesta análise.

Índice I de Moran Global

Considerando as seguintes hipóteses:

H_0 : Não há correlação espacial;
H_1 : Há correlação espacial.

Para os dados utilizados neste trabalho, usamos a função moran.test do pacote spdep em R, conforme Bivand e Wong (2018), com o propósito de obter um valor para o índice I de Moran. Obtivemos um resultado de 0,3628, um valor razoável que indica uma correlação positiva entre as regiões. A estimativa de erro padrão associada de 0,00011, que é menor do que 0,05, sugere uma evidência muito forte contra a hipótese nula de nenhuma correlação espacial nesses dados.

Índice I de Moran Local

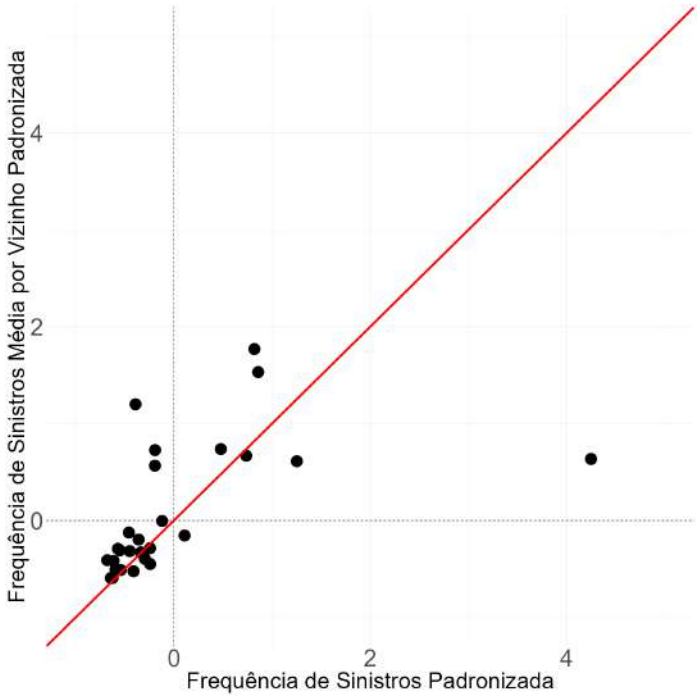

Figura 2.6: Diagrama de Disperção de Moran

Para análise da Figura 2.6, na qual a frequência de sinistros em cada estado e a frequência de sinistros média por vizinho foram padronizadas em torno do zero, nota-se que a grande maioria das regiões possui uma autocorrelação positiva, sendo 16 estados pertencentes ao quadrante com categoria baixo-baixo (frequência de sinistros e média dos vizinhos baixas) e 6 estados se encontram no quadrante com categoria alto-alto (frequência de sinistros e média dos vizinhos altas). Para as regiões com autocorrelação negativa, pode-se destacar 1 estado no quadrante alto-baixo (frequência de sinistros alta e média dos vizinhos baixa) e 3 estados no quadrante baixo-alto (frequência de sinistros baixa e média dos vizinhos alta). Em destaque, percebe-se a presenta de um outlier que possui uma frequência de sinistros muito alta, porém a média dos seus vizinhos, mesmo que positiva, é muito mais baixa. Este outlier é o estado de São Paulo, que como podemos ver, possui uma frequência de sinistros igual a 87.682 que, de fato, é um valor muito maior do que os valores dos seus vizinhos, sendo Minas Gerais com frequência igual 34.524, o segundo estado com maior frequência de sinistros. Com isso, conclui-se que a grande maioria dos estados possuem frequência de sinistros semelhante a média dos seus vizinhos.

O mapa de significância, Figura 2.7, mostra os locais com uma estatística local significativa, ou seja, com o grau de significância menor ou igual a 0,05, refletido em tons de

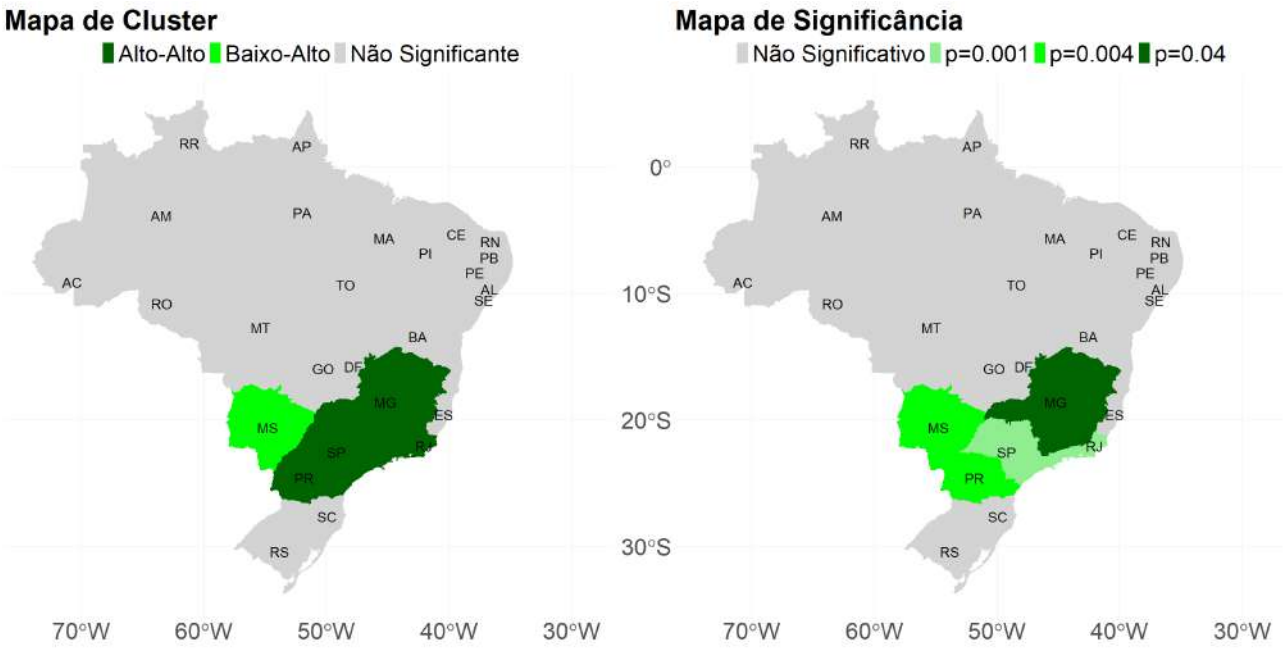

Figura 2.7: Mapas de Cluster e de Significância para o I de Moran Local

verde cada vez mais escuros. Apenas 4 estados possuem p-valores significativos, sendo o estado de Minas Gerais com o maior p-valor (0,04), São Paulo e Rio de Janeiro com os menores p-valores (0,001) e Mato Grosso do Sul e Paraná com p-valores intermediários (0,004).

O mapa de cluster, Figura 2.7, acrescenta às localizações significativas uma indicação do tipo de associação espacial em quatro categorias, como já explicado anteriormente. Neste trabalho, apenas duas categorias são representadas; foi utilizado o verde escuro para o cluster alto-alto (4 regiões) e o verde claro para o cluster baixo-alto (1 região). As demais regiões não possuem p-valor significativo. Foi considerado um nível de significância de 0,05, e portando 95% de confiança.

Portanto, podemos concluir que, por mais que no gráfico de dispersão, muitos estados aparentavam ter uma relação de frequência de sinistros parecida com as médias dos seus vizinhos, apenas os estados São Paulo, Rio de Janeiro, Minas Gerais, Paraná e Mato Grosso do Sul possuem uma autocorrelação que pode ser considerada significativa. Para São Paulo, Rio de Janeiro, Minas Gerais e Paraná a autocorrelação é positiva e os valores da frequência de sinistros são considerados altos e a média dos vizinhos também. Já o estado de Mato Grosso do Sul possui um autocorrelação negativa, onde a frequência de sinistros da região é considerada baixa se comparado a média dos vizinhos que é um

pouco mais alta.

Assim, podemos ver que por mais que o índice I de Moran global indique que existe correlação espacial em uma forma geral, quando olhamos especificamente para cada estado, podemos ver que a correlação espacial é significante, de fato, apenas para alguns estado nos dados utilizados.

Índice C de Geary

Considerando também, as mesmas hipóteses utilizadas para o índice I de Moran:

H_0 : Não há correlação espacial;
H_1 : Há correlação espacial.

Foi utilizada a função geary.test nos dados, também do pacote spdep em R, conforme Bivand e Wong (2018), para obter um valor para o índice C de Geary. O valor obtido foi de 0,6280, com uma estimativa de erro padrão associada de 0,0659. Novamente, o afastamento marcado da média de 1 indica forte correlação espacial positiva nos dados. Podemos rejeitar a hipóteses H_0, utilizando um nível de significância de 0,1, com 90% de significância.

Capítulo 3

Modelo Lineares Generalizados

No ramo estatístico, existem muitas situações na qual a variável resposta de interesse não é contínua, e por isso o modelo linear clássico com o componente aleatório seguindo uma distribuição normal não é o modelo mais adequado a ser utilizado. Diante disso, foi desenvolvido por Nelder e Wedderburn (1972), uma classe de modelos baseados na família exponencial com um parâmetro desconhecido, em que suas médias são não-lineares num conjunto de parâmetros lineares, são os chamado Modelos Lineares Generalizados.

Os Modelos Lineares Generalizados permitem, portanto, ampliar as suposições admitidas e examinar não somente as relações lineares entre as variáveis explicativas e a resposta, mas também analisar relações não lineares. Com eles, é possível modelar variáveis de interesse que assumem a forma de contagem, contínuas simétricas e assimétricas, binárias e categóricas. Uma das limitações dos GLM's é a exigência de que os erros sejam independentes.

Um Modelo Linear Generalizado consiste em três componentes: Componente Aleatório, o Componente Sistemático (Preditor Linear) e a Função de Ligação.

1. **Componente Aleatório**

   Neste componente é especificado a distribuição condicional da variável resposta Y_i dado os valores das variáveis explicativas do modelo para n observações amostradas independentes.

   No modelo de regressão linear temos que $Y_i \sim N(\mu_i, \sigma_i^2)$ e portanto a esperança de Y_i é definida como $E(Y_i) = \mu_i = \mathbf{X}\boldsymbol{\beta}$, em que o $\mathbf{X}$ representa o vetor de covariáveis e $\boldsymbol{\beta}$ representa o vetor de coeficientes do modelo.

   Porém, como já mencionado acima, interessados em situações mais genéricas, Nel-

der e Wedderburn (1972) propuseram uma situação onde a distribuição de Y_i é membro de uma família exponencial. Com isso a relação entre o valor esperado da variável resposta e as covariáveis pode retornar algo diferente do comum, nesse caso escrevemos a esperança como:

$$E(Y_i) = \mu_i = g(\mathbf{X}\boldsymbol{\beta}),$$

em que $\mathbf{X}$ representa o vetor de covariáveis, $\boldsymbol{\beta}$ representa o vetor de coeficientes do modelo e $g(.)$ é uma função genérica, chamada de função de ligação, que possui a missão de linearizar a relação entre a média e o preditor linear.

Abaixo vamos abordar sobre a família exponencial e posteriormente explicaremos com mais detalhes o papel da função de ligação.

Família Exponencial

Uma variável aleatória Y tem distribuição na família exponencial se a sua função de probabilidade ou função de densidade de probabilidade puder ser escrita como:

$$f(y) = c(y, \phi) exp\left\{ \frac{y\theta - a(\theta)}{\phi} \right\},$$

onde θ e ϕ são parâmetros canônico e de dispersão respectivamente e $a(\theta)$ e $c(y, \phi)$ são funções que determinam a função de probabilidade atual utilizada.

Para o $a(\theta)$, tem-se ainda, que:

$$E(y) = \dot{a}(\theta) \quad e \quad Var(y) = \phi\ddot{a}(\theta),$$

tais que $\dot{a}(\theta)$ e $\ddot{a}(\theta)$ representam a primeira e segunda derivada de $a(\theta)$ com respeito a θ, respectivamente.

Algumas das distribuições que fazem parte da família exponencial são: Binomial, Poisson, Gama, Gaussiana, Binomial Negativa, entre outras. A distribuição utilizada neste trabalho será a Poisson.

2. **Componente Sistemático**

Também conhecido como Preditor Linear, o Componente Sistemático é uma função linear de regressoras que pode ser escrito da seguinte forma:

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + ... + \beta_p X_{ip}.$$

Assim como no modelo linear, as regressoras X_{ij} são funções pré-especificadas das variáveis explicativas e, portanto, podem incluir variáveis explicativas quantitativas, transformações de variáveis explicativas quantitativas, regressores polinomiais, regressores dummy, interações e outras. Na verdade, uma das vantagens dos GLM's é que a estrutura do componente sistemático é familiar a de um modelo linear.

3. **Função de Ligação**

Como já mencionado, pode-se haver casos em que a relação entre as variáveis explicativas e variáveis respostas não são lineares. Por isso, a função de ligação possui o papel de linearizar a relação entre os componentes aleatórios (Valores observados de Y_i) e sistemáticos (vetor de coeficientes betas e variáveis explicativas X_i), onde essa função $g(.)$ é monótona e diferenciável, de tal forma que:

$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip}, \quad onde \quad \mu_i = E(Y_i).$$

Como a função de ligação é invertível, também podemos escrever:

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip}), \quad onde \quad \mu_i = E(Y_i).$$

Algumas Funções de Ligação		
Distribuição de Y	$g(\mu)$	Função de Ligação
Poisson	$log(\mu)$	Log
Normal	μ	Identidade
Binomial	$log(\frac{\mu}{1-\mu})$	Logito
Binomial	$\phi^{-1}(\mu)$	Probito
Binomial	$log(-log(1-\mu))$	Complemento Log-Log
Gama	μ^{-1}	Potência

Tabela 3.1: Tabela de Funções de Ligação.

A Tabela 3.1 mostra exemplos de funções de ligações que podem ser utilizadas na aplicação de Modelos Lineares Generalizados para algumas distribuições de Y, conforme Fox (2015).

Offset

Além da função de ligação, existe um outro fator muito importante, o offset.

Suponhamos o interesse em modelar o número de sinistros de automóveis em um grupo de risco, deve-se considerar o número de segurados expostos a este risco.

Denominamos a exposição ao risco como um vetor n_i. Se μ_i representa o vetor das médias dos números de sinistros Y_i, então o vetor da taxa de ocorrência θ_i pode ser calculado como μ_i/n_i, ou seja $\theta_i = \frac{\mu_i}{n_i}$. Com isso, pode-se escrever que:

$$g\left(\frac{\mu_i}{n_i}\right) = \mathbf{X}^T\boldsymbol{\beta},$$

em que $\mathbf{X}$ é o vetor de covariáveis e $\boldsymbol{\beta}$ é o vetor dos coeficientes do modelo.

Se a função de ligação for considerada como $g(.) = log(.)$, por exemplo, tem-se:

$$log\left(\frac{\mu_i}{n_i}\right) = \mathbf{X}^T\boldsymbol{\beta},$$

$$log(\mu_i) = log(n_i) + \mathbf{X}^T\boldsymbol{\beta}.$$

Neste caso, configura-se um modelo do tipo log-linear. Considerando o offset, pode-se dizer que Y_i possui um valor esperado diretamente proporcional a exposição, na qual:

$$e^{log(\mu_i)} = e^{log(n_i)+\mathbf{X}^T\boldsymbol{\beta}},$$

$$\mu_i = n_i e^{\mathbf{X}^T\boldsymbol{\beta}}.$$

Para estimar o modelo, foi utilizado inferência bayesiana pelo método de estimação MCMC, aplicando o pacote BRMS no RStudio, conforme R Core Team (2020), em que a função de verossimilhança utilizada foi a Poisson para todos os modelos e as distribuições à priori foram escolhidas pelo próprio pacote, sendo elas distribuições T-Student com diferentes parâmetros para cada estado, já que foram estimados modelos independente por estado, devido a autocorrelação espacial.

Capítulo 4

Análise dos Resultados

Como observado no capítulo 2, os dados utilizados possuem autocorrelação espacial. Por isso, decidimos aplicar o modelo GLM de forma independente para cada estado e assim poder observar se a frequência de sinistros de cada estado é parecida com os estados próximos ou se possuem alguma relação.

Inicialmente, foi testado um modelo que considerasse três variáveis aleatórias, sexo, faixa etária e categoria do carro. Porém a variável categoria do carro não apresentava indícios de ser significante para explicar a frequência de sinistros para os dados utilizados, já que quando ela estava no modelo não era possível verificarmos as diferenças de ocorrências de sinistros para as diversas características, os resultados ficavam muito semelhantes. Portanto, optamos por retirá-la do modelo. Isso pode ter ocorrido pela razão de o dado possuir pouquíssimos casos de sinistros ocorridos com carros importados, fato que interferia diretamente no modelo fazendo com que obtivéssimos resultados irrelevantes.

Dito isso, definimos Y_{ij} uma variável aleatória que representa o número de sinistros da i-ésima observação para um estado brasileiro j. Considere também duas covariáveis X_{1j} (sexo) e X_{2j} (faixa etária), definidas da seguinte forma:

$$X_{1j} = \begin{cases} 0, Feminino; \\ 1, Masculino. \end{cases}$$

$$X_{2j} = \begin{cases} 0, Acima\ de\ 35\ anos; \\ 1, Entre\ 18\ e\ 35\ anos. \end{cases}$$

Suponha que:

$$Y_{ij}|\theta_{ij} \sim Poisson(\lambda_{ij}),$$

de tal forma que $\lambda_{ij} = E_{ij}\theta_{ij}$ e portanto obtemos uma taxa de ocorrência definida como $\theta_{ij} = \frac{\lambda_{ij}}{E_{ij}}$, onde E_{ij} é a exposição ao risco na i-ésima observação para o estado j e θ_{ij} é a probabilidade de ocorrência de sinistro na i-ésima observação para o estado j.

Em vista disso, queremos explicar, em função das regressoras X_{ij}, a ocorrência de sinistros Y_{ij} dado uma exposição ao risco θ_{ij}, conforme descrito abaixo:

$$E(Y_{ij}|\theta_{ij}) = \mathbf{X}\boldsymbol{\beta},$$

em que $\mathbf{X}$ é o vetor de covariáveis composto pelos componentes destas covariáveis X_{ij} e $\boldsymbol{\beta}$ é o vetor dos coeficientes do modelo relacionado as covariáveis.

Finalmente, assumindo a função de ligação log, e com isso, o modelo para cada estado j será dado da seguinte forma:

$$log\left(\frac{\lambda_{ij}}{E_{ij}}\right) = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij},$$

$$log(\lambda_{ij}) = log(E_{ij}) + \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij}.$$

Voltando para a escala original, tem-se:

$$exp(log(\lambda_{ij})) = exp(log(E_{ij}) + \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij}),$$

$$\lambda_{ij} = E_{ij}e^{\beta_{0j}+\beta_{1j}X_{1ij}+\beta_{2j}X_{2ij}}.$$

Quando fazemos o produto de E_{ij} com θ_{ij}, obtemos um número médio de sinistros. Porém, podemos estar interessados em olhar apenas para o θ_{ij}, que nos proporciona uma taxa média de ocorrência de sinistros, desta forma temos:

$$\theta_{ij} = e^{\beta_{0j}+\beta_{1j}X_{1ij}+\beta_{2j}X_{2ij}}.$$

Dito isso, utilizando a função de distribuição Poisson, aplicamos o modelos com o auxílio do pacote BRMS no RStudio, conforme R Core Team (2020). Foram feitas duas cadeias, com o objetivo de obter um resultado mais confiável, com 10.000 iterações em cada; destas 10.000 iterações, 2.000 foram de warmup (aquecimento) e thin igual a 1,

com isso verificou-se que o modelo não possui autocorrelação. As duas cadeias do MCMC convergiram e abaixo seguem alguns resultados.

Em um primeiro momento, foram construídos gráficos com intervalos de 95% de credibilidade para os coeficientes β_{ij}. Com isso, é possível verificar se as covariáveis são significativas para explicar a ocorrência de sinistros em cada estado.

Ao observar as Figuras 4.1, 4.2 e 4.3, podemos notar que em relação ao intercepto, para todos os estados obtivemos um resultado significante, visto que nenhum dos intervalos incluem o zero, porém isso não é uma realidade para β_1 e para β_2.

Para a covariável sexo (Figura 4.2), o gráfico nos mostra que nos estados Roraima, Rondônia, Pará, Mato Grosso do Sul, Maranhão, Espírito Santo e Alagoas, o sexo não é significante para explicar a ocorrência de sinistros. Já na Figura 4.3, os estados onde a variável faixa etária não é significativa para explicar a ocorrência de sinistros são: Santa Catarina, Rio Grande do Sul, Roraima, Rondônia, Pernambuco, Maranhão, Amapá, Amazonas, Alagoas e Acre. Para os demais estados, as variáveis são significativas.

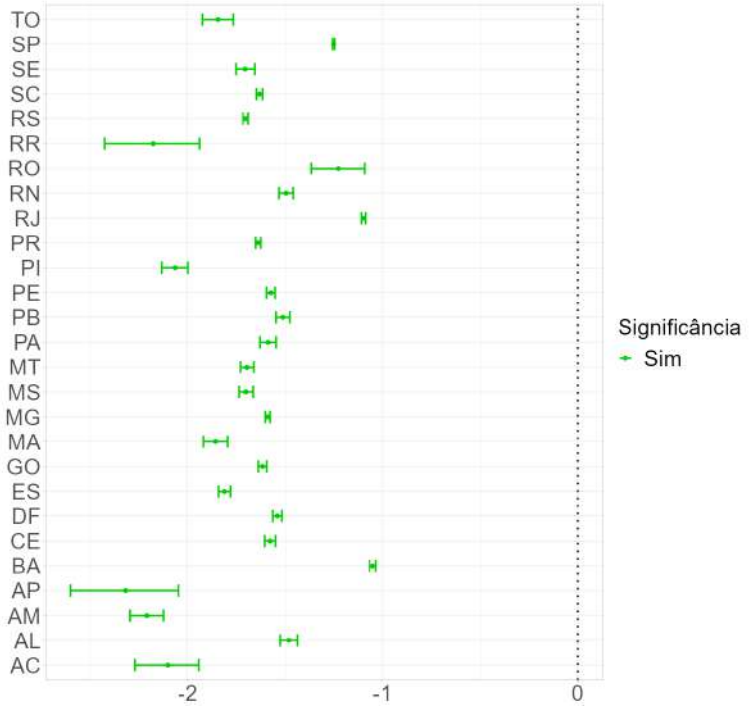

Figura 4.1: Gráfico de Intervalo de 95% de Credibilidade para β_0 (intercepto).

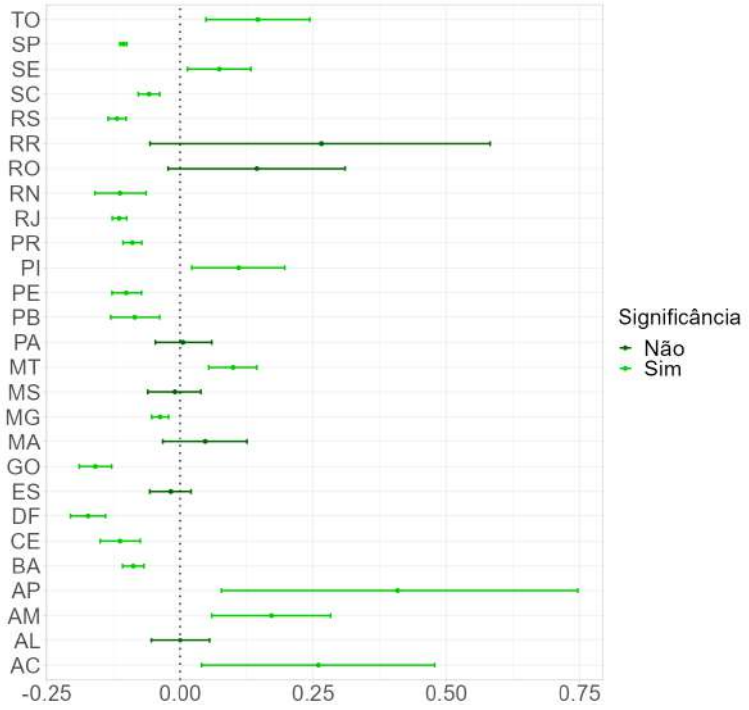

Figura 4.2: Gráfico de Intervalo de 95% de Credibilidade para β_1 (sexo).

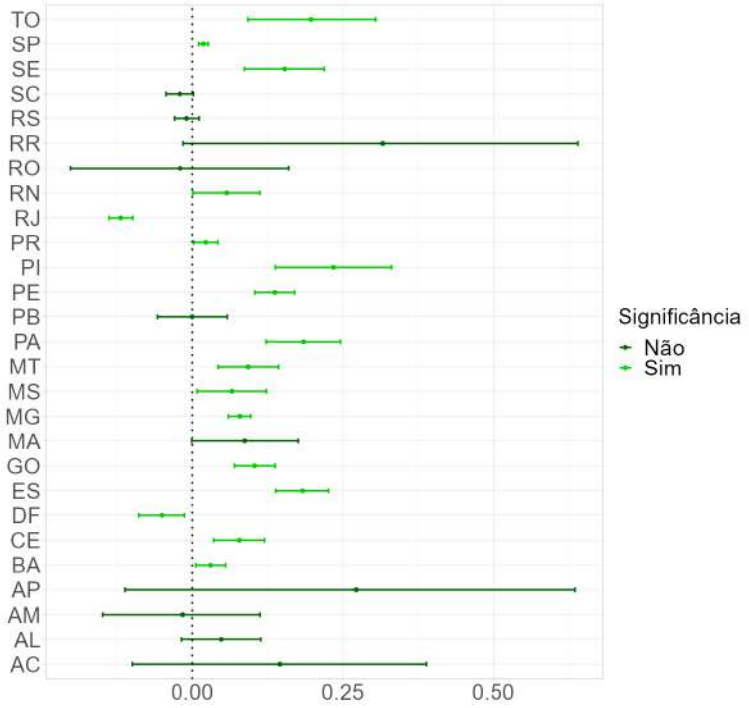

Figura 4.3: Gráfico de Intervalo de 95% de Credibilidade para β_2 (faixa etária).

Um fator em destaque é a amplitude dos intervalos de alguns estados, como o Amapá, que possui o maior intervalo de credibilidade, resultado da baixa frequência relativa de sinistros (290), que em comparação aos demais estados, é o que possui menor frequência relativa, seguido de Roraima (937), Acre (1.239) e Rondônia (1.425). Nestes casos, pelo tamanho da amostra ser bem pequena, obtemos uma estimativa menos precisa. Já no estado de São Paulo, ocorre o oposto, o intervalo é o menor existente, pois é o estado com maior frequência relativa de sinistros (87.682), seguido de Minas Gerais (34.524), Paraná (27.546) e Rio de Janeiro (26.858). Dessa forma, obtém-se estimativas com maior precisão para estes estados. Essa é uma adversidade que acontece devido a base de dados utilizada possuir pouca informação.

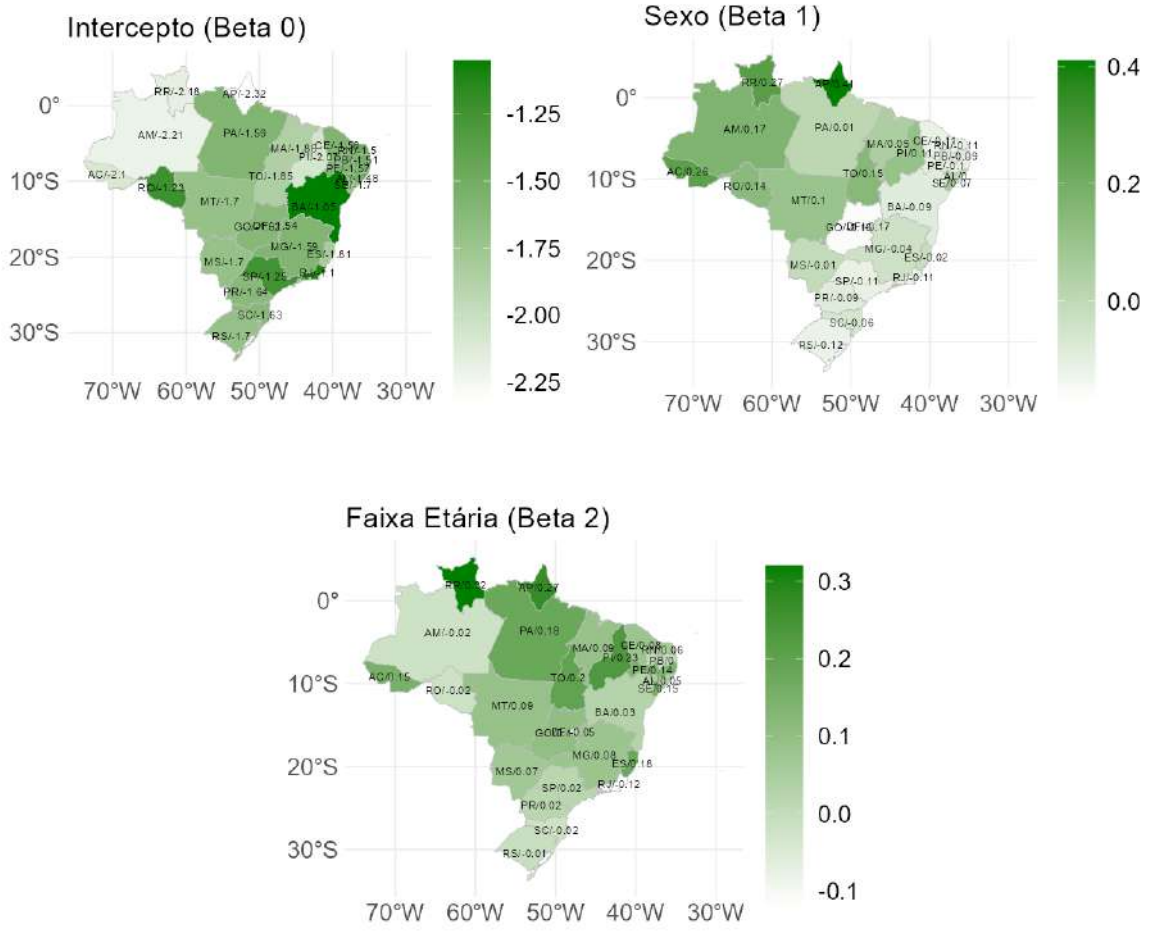

Figura 4.4: Mapas para os valores de β_{ij}.

Uma segunda análise possível é relacionada ao mapa da Figura 4.4, que nos mostra os valores dos coeficientes β_{ij} para cada estado. A escala dos mapas está diferente para melhor visualização dos efeitos deste coeficiente, já que quando colocado em uma mesma

escala não é possível observar a diferença entre os estados j.

Nestes mapas, podemos verificar que, para o β_{1j} as regiões Norte e parte do Centro-Oeste possuem valores que indicam um efeito positivo da covariável sexo na probabilidade de ocorrência de sinistros. Já quando olhamos para as regiões Nordeste, Sudeste e Sul percebemos um efeito negativo da covariável sexo.

Em relação ao β_{2j}, nota-se que a grande maioria dos estados das regiões Norte, Nordeste, Centro-Oeste e Sudeste possuem um efeito positivo da variável explicativa faixa-etária na probabilidade de frequência de sinistros, contendo apenas algumas excessões como Rondônia, Distrito Federal e Rio de Janeiro. Enquanto no Sul tem-se a maioria dos estados com um efeito negativo sobre a variável resposta, sendo apenas o Paraná com efeito positivo.

Por fim, para β_{0j}, todos os valores indicam efeito negativo, em que a região Norte possui valores mais afastados de zero, indicando efeito negativo maior.

No geral, podemos verificar que as regiões Norte e parte do Centro-Oeste possuem valores para os β_{ij} maiores do que as outras regiões, indicando um efeito maior das covariáveis na probabilidade de ocorrência de sinistros. Isso pode acontecer devido ao número de frequência de sinistros nessas áreas, que é menor do que nas demais.

A terceira análise realizada está relacionada com a estimativa da taxa de ocorrência θ_{ij} para cada estado j, abaixo mostrada através de uma mapa de calor, considerando os quatro cenários possíveis.

Ao explorar o mapa (Figura 4.5) abaixo, é notório que, no geral, pessoas do sexo feminino possuem uma taxa de ocorrência de sinistros maior do que pessoas do sexo masculino. Já para a covariável idade, as taxas de ocorrência parecem ser bem similares para ambas as faixas etárias, no entando, se observarmos com mais atenção seria possível notar uma leve diferença, revelando que segurados com idade entre 18 e 35 anos possuem uma taxa de ocorrência maior do que segurados acima de 35 anos. Nota-se ainda, que a taxa de ocorrência dos estados são similares as taxas dos estados vizinhos, exceto alguns casos um pouco destoantes, como Rondônia e Bahia. Então, de fato, existe uma autocorrelação espacial nos dados. Percebe-se, que no geral, estados de uma mesma região possuem taxas de ocorrência mais similares, como podemos ver Rio de Janeiro e São Paulo; Minas Gerais e Espírito Santo; Paraná, Santa Catarina e Rio Grande do Sul; Mato Grosso e Pará; no Nordeste podemos perceber isso também.

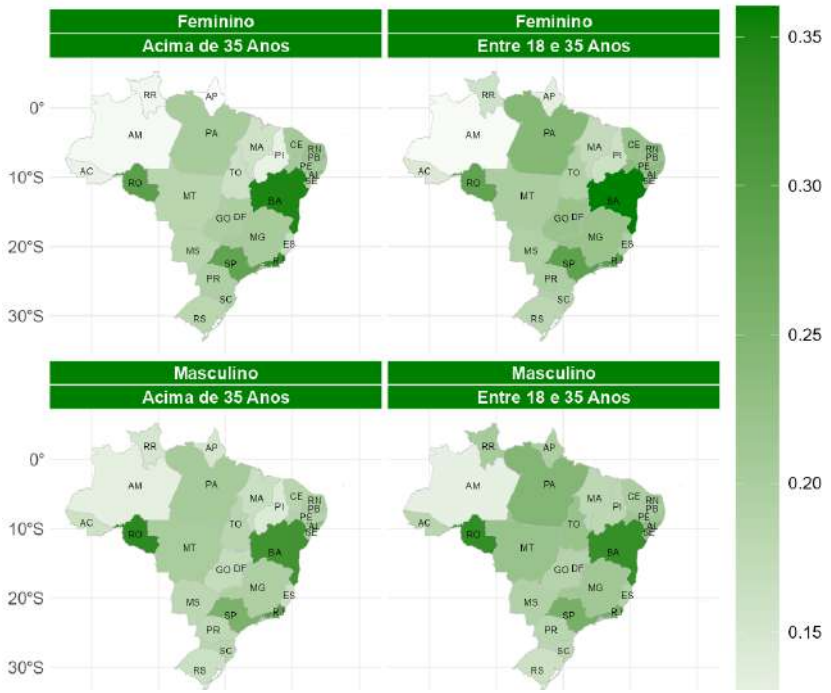

Figura 4.5: Mapa com a estimativa para θ_{ij}, taxa média de ocorrência de sinistros para cada estado j.

Para uma avaliação do comportamento entre as taxas de ocorrência θ_{ij} de cada estado j, podemos fazer uma razão delas para estados diferentes, considerando um mesmo cenário. Seja o cenário sexo feminino ($x_{i1j} = 0$) e faixa etária entre 18 e 35 anos ($x_{i2j} = 1$), estimamos os valores de θ_{ij} para cada estado j no RStudio e obtemos os seguintes resultados para alguns estados: São Paulo ($\theta_{iSP} = 0,2912$), Rio de Janeiro ($\theta_{iRJ} = 0,2954$), Paraná ($\theta_{iPR} = 0,1987$) e Santa Catarina ($\theta_{iSC} = 0,1915$). Com isso podemos realizar uma razão entre as taxas.

Para **Rio de Janeiro** e **São Paulo**: $\frac{\theta_{iRJ}}{\theta_{iSP}} = \frac{0,2954}{0,2912} = 1,0144$.

Logo, conluí-se que no Rio de Janeiro a taxa de ocorrência é maior em 1,44%, o que é uma porcentagem bem pequena, mostrando que, de fato, a taxa de ocorrência de sinistros entre Rio de Janeiro e São Paulo é bem parecida.

Para **São Paulo** e **Paraná**: $\frac{\theta_{iSP}}{\theta_{iPR}} = \frac{0,2912}{0,1987} = 1,4656$.

Já para São Paulo e Paraná, possui uma diferença maior, em que São Paulo obtém uma taxa de ocorrência de sinistros 46,56% maior do que o Paraná, quase o dobro. O

mesmo ocorre entre Rio de Janeiro e Paraná.

Para **Rio de Janeiro** e **Paraná**: $\frac{\theta_{iRJ}}{\theta_{iPR}} = \frac{0{,}2954}{0{,}1987} = 1,4868$.

Porém, se olharmos para uma razão de taxas entre Paraná e Santa Catarina, conforme abaixo, podemos ver que Paraná possui uma ocorrência de sinistros 3,74% maior do que Santa Catarina, o que também é uma diferença muito pequena. O mesmo ocorreria se comparássemos Ceará com Rio Grande do Norte (Nordeste), ou Mato Grosso e Mato Grosso do Sul (Centro-Oeste), isto é, estados da mesma região possuem taxa de ocorrência de sinistros similares, sendo esta mais uma análise que confirma o fato de haver autocorrelação espacial nos dados.

Para **Paraná** e **Santa Catarina**: $\frac{\theta_{iPR}}{\theta_{iSC}} = \frac{0{,}1987}{0{,}1915} = 1,0374$.

Como uma análise adicional, podemos realizar uma razão de taxas entre os diferentes cenários para alguns estados, e assim obter conclusões sobre as covariáveis em relação a taxa de ocorrência de sinistros. Dado o nosso modelo já descrito acima, temos:

$$\theta_{ij} = e^{\beta_{0j}+\beta_{1j}X_{1ij}+\beta_{2j}X_{2ij}}.$$

Se considerarmos cenários diferentes, em que:

- **Cenário 1:** Sexo Masculino e Faixa Etária Acima de 35 Anos;
- **Cenário 2:** Sexo Feminino e Faixa Etária Acima de 35 Anos.

Se definirmos a razão de taxas para diferentes cenários como RT, podemos escrevê-la como:

$$RT = \frac{Cenário\ 1}{Cenário\ 2} = \frac{E(Y_{1j}|X_{1j}=1, X_{2j}=0)}{E(Y_{1j}|X_{1j}=0, X_{2j}=0)} = \frac{e^{\beta_{0j}+\beta_{1j}X_{1j}}}{e^{\beta_{0j}}} = e^{\beta_{1j}X_{1j}}.$$

Ou seja, se pensarmos em comparar a variável sexo temos que a razão de taxas é $RT = e^{\beta_{1j}X_{1j}}$. Da mesma forma, se compararmos a variável faixa etária, a razão de taxas pode ser escrita como: $RT = e^{\beta_{2j}X_{2j}}$.

Com isso, foram escolhidos três estados diferentes, em que todos os três consideraram as covaráveis significativas para explicar a ocorrência de sinistros, para realizar a razão de taxas: São Paulo, Rio de Janeiro e Paraná.

São Paulo:

Em relação a São Paulo, tem-se as seguintes estimativas para β_{ij}:

$$\theta_{iSP} = e^{-1,2522-0,1063X_{i1SP}+0,0184X_{i2SP}}.$$

Se focarmos na variável sexo em que $X_1 = 1$ é masculino, temos: $e^{\beta_1} = e^{-0,1063} = 0,2859$. Ou seja, olhando para o estado de São Paulo, pessoas do sexo feminino cometem sinistros 28,59% mais do que pessoas do sexo masculino.

Se direcionarmos a atenção para a variável faixa etária, onde $X_2 = 1$ é entre 18 e 35 anos, tem-se: $e^{\beta_2} = e^{0,0184} = 1,0185$. Portanto, no estado de São Paulo, indivíduos com idade entre 18 e 35 anos cometem 1,85% mais sinistros do que pessoas acima de 35 anos.

Rio de Janeiro:

Em relação ao Rio de Janeiro, tem-se as seguintes estimativas para β_{ij}:

$$\theta_{iRJ} = e^{-1,1006-0,1142X_{i1RJ}-0,1189X_{i2RJ}}.$$

Para a variável sexo em que $X_1 = 1$ é masculino, obtemos: $e^{\beta_1} = e^{-0,1142} = 0,8921$. Ou seja, olhando para o estado do Rio de Janeiro, pessoas do sexo feminino cometem sinistros 89,21% mais do que pessoas do sexo masculino.

Já para a variável faixa etária, onde $X_2 = 1$ é entre 18 e 35 anos, tem-se: $e^{\beta_2} = e^{-0,1189} = 0,8879$. Portanto, no estado do Rio de Janeiro, indivíduos com idade acima de 35 anos cometem 88,79% mais sinistros do que pessoas entre 18 e 35 anos.

Paraná:

Em relação ao Paraná, tem-se as seguintes estimativas para β_{ij}:

$$\theta_{iPR} = e^{-1,6385-0,0895X_{i1PR}+0,0224X_{i2PR}}.$$

Relacionado a variável sexo, onde $X_1 = 1$ é masculino, obtemos: $e^{\beta_1} = e^{-0,0895} = 0,9143$. Ou seja, olhando para o estado do Paraná, pessoas do sexo feminino cometem sinistros 91,43% mais do que pessoas do sexo masculino.

Já para a variável faixa etária, onde $X_2 = 1$ é entre 18 e 35 anos, tem-se: $e^{\beta_2} = e^{0,0224} = 1,0227$. Portanto, no estado do Paraná, indivíduos com idade entre 18 e 35 anos cometem 2,27% mais sinistros do que pessoas acima de 35 anos.

Capítulo 5

Conclusão

Nesse trabalho, procurou-se propor a utilização de Modelos Lineares Generalizados com função de distribuição Poisson para o ajuste de frequência de sinistros de automóveis, considerando idade e sexo do condutor para os diferentes estados do Brasil, tendo como dois objetivos explicar a frequência de sinistros utilizando as covariáveis sexo e faixa etária e verificar o comportamento das regiões em relação a frequência de sinistros.

Para estimarmos o modelo, utilizamos a distribuição Poisson, aplicada a Modelos Lineares Generalizados. De acordo com os resultados obtidos nos ajustes, as características do segurado podem impactar na frequência de sinistros. Para a variável explicativa sexo, obtivemos que, mulheres cometem mais sinistros do que homens. Já para a variável explicativa faixa etária, foi notório que, geralmente indivíduos entre 18 e 35 anos possuem maior frequência de sinistros do que pessoas acima de 35 anos, com algumas excessões como Rio de Janeiro, por exemplo.

Em relação aos gráficos de intervalo de 95% de credibilidade, percebemos que para as covariáveis sexo e faixa etária, alguns estados não as consideraram significativas para explicar a frequência de sinistros; estes são os estados que possuem o zero incluso nos seus intervalos. Já para os que não possuem o zero, podemos concluir que consideraram as covariáveis significativas para explicar a variável resposta.

Um ponto importante sobre os intervalos de credibilidade citado no capútulo 4, foi a amplitude de alguns intervalos, que por serem muito grandes, fazem com que as nossas estimativas tenham uma precisão menor. Uma solução eficaz para este problema, seria uma tentativa de agrupar os estados que são próximos e que possuem baixa ocorrência de sinistros. Por exemplo, pode-se citar Roraima e Amazonas, são estados que possuem fronteira e baixa frequência de sinistros, ou então Amapá e Pará, que se encontram na

mesma situação. Posto isso, seria possível uma análise utilizando grupos de estados que teriam uma frequência de sinistros maior, nos proporcionando um melhor resultado.

Uma outra conclusão que podemos ressaltar é que estados próximos, de fato, possuem uma taxa de ocorrência de sinistros semelhantes devido a autocorrelação espacial. Como foi observado no Capítulo 4, estados da mesma região possuem frequência de sinistros muito parecidas, como por exemplo Rio de Janeiro e São Paulo (diferença de 1,44%) e Paraná e Santa Catarina (diferença de 3,74%). Dito isso, uma possibilidade interessante seria pensar em uma análise por região, o que também poderia nos trazer melhores resultados.

Mesmo diante de fatores que tornam a análise feita neste trabalho muito relevante, é possível notar pontos de atenção que podem ser melhorados. Um desses pontos é que a distribuição Poisson considera média e variância iguais, o que caracteriza uma equidispersão, que nem sempre é observado em dados de contagem. Nos dados utilizados, por exemplo, podemos observar uma sobredispersão nos dados, que é definida pela variância maior do que a média (testado no RStudio).

Diante disso, pode-se pensar em muitas distribuições baseadas na de Poisson, porém que possuem a capacidade de lidar com a sobredispersão dos dados, segundo Pereira (2016).

Em geral, quando a variância é diferente da média, temos algumas alternativas, como a distribuição Binomial Negativa ou a distribuição de Poisson Generalizada, que possuem um parâmetro adicional, tornando-se mais flexíveis. Entretanto, segundo Carvalho (2021), devido a este parâmetro adicional, essas distribuições são mais susceptíveis a erros na estimação desses parâmetros, portanto, como alternativa à distribuição de Poisson e com apenas um parâmetro, essa tese sugeriu a distribuição Bell, que de acordo com Castellares et al. (2018), a distribuição Bell é uni-paramétrica e pode ser aplicada a variáveis respostas que se tratam de contagem, sendo possível a aplicação de um modelo de regressão relacionado a um preditor linear por meio de uma função de ligação, na mesma configuração que um GLM. Este modelo possui funções que são consideradas de fácil manuseio e é uma boa alternativa para utilizar no lugar da distribuição Poisson, sendo ele uma possibilidade para trabalhos futuros.

Um segundo ponto de atenção está relacionado as limitações dos GLM's, que exige que os erros sejam independentes. Isso pode significar uma certa dificuldade de modelar bancos de dados com estruturas longitudinais, espaciais ou multiníveis. Como os dados utilizados possuem uma estrutura espacial, considerando frequência de sinistros

por estado, é grande a chance do pressuposto de independência ser violado, sendo este, um fator de impedimento para um melhor resultado. Porém, é possível contornar essa situação utilizando Modelos Lineares Generalizados Mistos ou Equações de Estimações Generalizadas, exemplos de modelos que poderiam ser utilizados também em um próximo trabalho. Ou então, se pensarmos em algo um pouco mais elaborado, seria possível a aplicação de um modelo chamado CAR, que é uma classe de modelos que, segundo Monteiro et al. (2004), incorpora a autocorrelação espacial entre as observações.

Referências Bibliográficas

Anselin, L. (1995) Local indicators of spatial association—lisa. *Geographical analysis*, **27**, 93–115.

AUTOSEG (2020) *AUTOSEG - SISTEMA DE ESTATÍSTICAS DE AUTOMOVEIS DA SUSEP*. URLhttps://www2.susep.gov.br/menuestatistica/Autoseg/principal.aspx.

Banerjee, S., Carlin, B. P. e Gelfand, A. E. (2014) *Hierarchical modeling and analysis for spatial data*. CRC press.

Bivand, R. e Wong, D. W. S. (2018) Comparing implementations of global and local indicators of spatial association. *TEST*, **27**, 716–748.

Bürkner, P. () An r package for bayesian multilevel models using stan. *J Statist Software*.

Carvalho, R. M. d. (2021) Modelagem da frequência de sinistros de automóveis nas regiões sul e sudeste do brasil: um estudo de caso utilizando a distribuição bell.

Castellares, F., Ferrari, S. L. e Lemonte, A. J. (2018) On the bell distribution and its associated regression model for count data. *Applied Mathematical Modelling*, **56**, 172–185.

CNseg (2020) *CNseg - A Confederação Nacional das Empresas de Seguros Gerais, Previdência Privada e Vida, Saude Suplementar e Capitalização*. URLhttps://cnseg.org.br/.

Dobson, A. J. e Barnett, A. G. (2018) *An introduction to generalized linear models*. CRC press.

Ferreira, J. L. M. (2013) Modelos de regressão para a previsão de sinistros.

Fox, J. (2015) *Applied regression analysis and generalized linear models*. Sage Publications.

IBGE (2020) *IBGE - Instituto Brasileiro de Geografia e Estatistica*. URLhttps://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/28668-ibge-divulga-estimativa-da-populacao-dos-municipios-para-2020.

IBPAD (2020) *IBPAD*. URLhttps://ibpad.com.br/ciencia-dados/o-que-e-estatistica-bayesiana/.

Monteiro, A. M. V., Câmara, G., Carvalho, M. e Druck, S. (2004) Análise espacial de dados geográficos. *Brasília: Embrapa*.

Nelder, J. A. e Wedderburn, R. W. (1972) Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, **135**, 370–384.

Oliveira, C. A. d. (2008) Análise espacial da criminalidade no rio grande do sul.

PAIVA, C. (2007) Dependência espacial. *Setores censitários, Zonas OD, Distritos, Prefeituras etc... CET/SP e PUC/SP*.

Pereira, H. T. (2016) Estudo da distribuição de poisson generalizada.

R Core Team (2020) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URLhttps://www.R-project.org/.

SUSEP (2020) *SUSEP - Superintendência de Seguros Privados*. URLhttps://www.gov.br/susep/pt-br.