Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Análise Espacial dos Acidentes nas Rodovias Federais do Estado de Goiás e do Distrito Federal.

Gabriel Alvares de Faria - 14/0139818

Brasília

Dezembro 2018

Gabriel Alvares de Faria **14/0139818**

Análise Espacial dos Acidentes nas Rodovias Federais do Estado de Goiás e do Distrito Federal.

Relatório apresentado à disciplina Estágio Supervisionado II do curso de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientador: Prof. Dr. Alan Ricardo da Silva

Brasília

2018

Resumo

Este trabalho tem por objetivo analisar acidentes de trânsito e seus possíveis fatores influentes a partir de técnicas de estatística espacial. Para tanto, utilizou-se uma base de dados da Polícia Rodoviária Federal na qual os acidentes são georreferenciados.

Assim, optou-se pela regressão logística geograficamente ponderada (RLGP), um tipo modelo espacial local, para se analisar os acidentes de trânsito nas rodovias federais do estado de Goiás e do Distrito Federal. Explorou-se as possibilidades ofertadas pela estatística espacial, como a utilização de mapas, para apresentar a distribuição espacial dos parâmetros, assim como as probabilidades estimadas.

O uso de um modelo espacial local se mostrou mais eficiente do que um modelo clássico. No entanto, o modelo proposto para explicar a ocorrências de acidentes com feridos se mostrou inadequado, com uma baixa variabilidade explicada e sobrevalorizando as probabilidades estimadas.

Palavras-chaves: Regressão espacial, Acidentes de trânsito, RGP, RLGP, Regressão Logística

Lista de Tabelas

Lista de Figuras

Sumário

Capítulo 1

INTRODUÇÃO

O principal meio de transporte para a movimentação de cargas e de passageiros no Brasil é o modo rodoviário (CNT, 2017). O modo rodoviário como meio de transporte e escoamento de cargas e passageiros se intensificou a partir do governo Juscelino Kubitschek, com a construção de rodovias e por políticas de incentivo à indústria automobilística (SCHROEDER e CASTRO, 1996). De fato, segundo a CNT (2017), tem-se que o modo rodoviário é responsável por cerca de 61% do transporte de cargas. Em 2017, o Brasil detinha uma frota circulante de 97.091.956 veículos (DENATRAN, 2017). E, em relação à infraestrutura existente, a malha rodoviária brasileira possui uma extensão total de 1.578.102 quilômetros, sendo 76.259 quilômetros de rodovias federais e o restante de rodovias estaduais e municipais (CNT, 2017).

Nos mais de 76 mil quilômetros de rodovias federais ocorreram, segundo balanço da PRF (2017), 89.318 acidentes com 6.244 mortes e 83.978 feridos em 2017. Ainda segundo este balanço, foram 5.853.185 autos de infração emitidos pela PRF em 2017, sendo mais da metade por excesso de velocidade e 224.479 por ultrapassagem proibida. Um outro dado importante, retirado da base de dados, acessível para download, da PRF sobre os acidentes rodoviários, é que, desses 89.318 acidentes, aproximada-

mente 15% deles ocorreram sob chuva.

Tendo em vista as informações descritas acima, esse trabalho tem por objetivo analisar os acidentes de trânsito em rodovias federais do estado de Goiás e do Distrito Federal no ano de 2017 e seus possíveis fatores influentes utilizando uma abordagem espacial. Para tanto, será utilizada a técnica de estatística espacial de área para relacionar os acidentes rodoviários observados com as condições climáticas e aspectos físicos de rodovia que podem ter influenciado na ocorrência dos mesmos.

1.1 OBJETIVOS

O objetivo geral do trabalho é analisar os acidentes nas rodovias federais do estado de Goiás e do Distrito Federal utilizando técnicas de estatística espacial.

Os objetivos específicos são:

- Mapear os acidentes com feridos nos estados de Goiás e no Distrito Federal.
- Identificar as rodovias e trechos mais perigosos.
- Identificar variáveis que afetam a ocorrência dos acidentes com feridos.

Capítulo 2

ESTATÍSTICA ESPACIAL

2.1 Introdução

A estatística espacial pode ser resumida no que Tobler (1970) definiu como sendo a Primeira Lei da Geografia: "todas as coisas são parecidas, mas coisas mais próximas se parecem mais do que coisas mais distantes". Em outras palavras, a estatística espacial ignora o postulado clássico de observações independentes, possibilitando que exista um determinado tipo de dependência nos dados.

Enxerga-se dados espaciais como uma realização de um processo estocástico. Assim, existe um mecanismo gerador $\{Z(u), u \epsilon A\}$ responsável pelos dados observados, em que Z é o atributo analisado e u é um ponto qualquer pertencente à região de estudo A (Druck *et al.*, 2004).

Existem três tipos de modelagem em estatística espacial: processo pontual, análise de área e análise de superfície. Para este trabalho, optou-se por empregar a análise de área, com uma pequena aplicação de análise de superfície.

2.2 Análise de área

A análise de área é empregada quando o fenômeno em estudo é expresso por dados agregados e delimitados por diferentes áreas que compõem a região em estudo (Druck *et al.*, 2004). Sua aplicação é comumente utilizada para ajustar modelos de regressão que, considerando a estrutura espacial dos dados, consigam explicar o padrão de ocorrência do fenômeno de interesse (Druck *et al.*, 2004).

Ao levar em consideração a localização espacial do fenômeno em estudo de forma explícita, os modelos espaciais conseguem mexer na estrutura dos dados de modo a remover a dependência espacial dos resíduos.

Existem duas classes de modelos na estatística espacial por área: os modelos com efeitos globais e os com efeitos locais (Druck *et al.*, 2004). Em geral, para uma análise de regressão em uma mesma base de dados, tem-se que:

$$R^2_{clássico} \leq R^2_{global} \leq R^2_{local},$$

onde R^2 é o coeficiente de determinação do modelo (Fotheringham *et al.*, 2002).

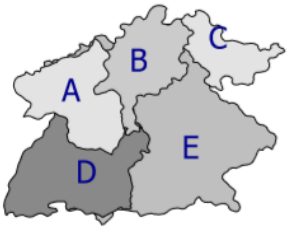
Apenas a modelagem local será abordada nesse trabalho, pois diferentemente da abordagem global que assume o processo gerador dos dados como sendo estacionário, os modelos locais permitem que haja diferentes padrões de associação espacial entre as variáveis do modelos, $i.e.$, os parâmetros dos modelos locais podem variar no espaço devido a uma heterogeneidade espacial do fenômeno em estudo (Fotheringham *et al.*, 2002).

2.2.1 Dependência espacial

Uma análise exploratória é requerida para a elaboração do melhor modelo. Além dos mapas que fornecem uma primeira impressão sobre o fenômeno em estudo, analisa-se a estrutura de dependência espacial dos dados. O índice global de Moran (*I de Moran*) fornece uma magnitude geral da correlação espacial do conjunto de observações e é dado por (Druck *et al.*, 2004)

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(z_i - \overline{z})(z_j - \overline{z})}{\sum_{i=1}^{n} (z_i - \overline{z})^2}, \tag{2.1}$$

onde n é o número de áreas na região de estudo, $\mathbf{W}$ é a matriz de vizinhança normalizada pela linha e que indica se as áreas i e j são vizinhas ou não, z_i o valor atribuído à área i e $\overline{z}$ é a média da região de estudo. Um exemplo da matriz $\mathbf{W}$ de vizinhança pode ser vista na Figura 2.1. Valores do índice *I de Moran* próximos a 0 indicam independência espacial dos dados e valores que em módulo são próximos a 1 indicam uma associação espacial.



	A	B	C	D	E
A	0	0,5	0	0,5	0
B	0,25	0	0,25	0,25	0,25
C	0	0,5	0	0	0,5
D	0,33	0,33	0	0	0,33
E	0	0,33	0,33	0,33	0

Figura 2.1: Exemplo de uma matriz W de vizinhança.
Fonte:Druck *et al.* (2004)

Note que é possível reescrever o índice *I de Moran* na forma

$$\mathbf{I} = \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{W}\mathbf{Z}, \tag{2.2}$$

que corresponde ao parâmetro estimado da regressão linear de **WZ** em **Z**, onde **Z** é a matriz dos valores normalizados das áreas e, **WZ** é a matriz das médias dos valores normalizados dos vizinhos de cada área. De fato, uma outra maneira de analisar a autocorrelação amostral é por meio do diagrama de espalhamento de Moran, que traz toda essa informação na forma de um gráfico (Anselin, 1996). Tal gráfico é exibido na Figura 2.2.

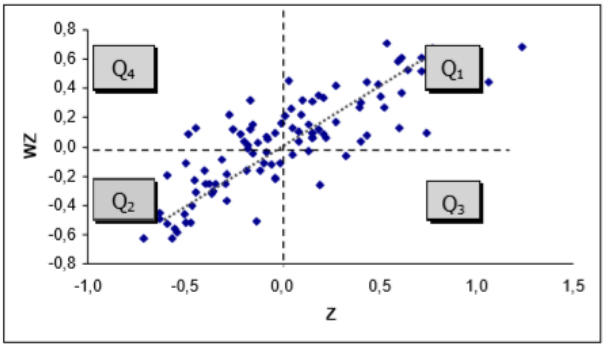


Figura 2.2: Diagrama de espalhamento de Moran.
Fonte:Druck *et al.* (2004)

Para verificar se essa estrutura de dependência espacial é a mesma em todo o espaço, *i.e.*, verificar se existe heterogeneidade espacial no fenômeno em estudo, calcula-se o índice local de Moran para cada área i que é dado por (Druck *et al.*, 2004)

$$I_i = \frac{z_i \sum_{j=1}^{n} w_{ij} z_j}{\sum_{j=1}^{n} {z_j}^2}, \tag{2.3}$$

onde z_i é o valor normalizado atribuído à área i, n é o número de áreas na região de estudo e **W** é a matriz de vizinhança que indica se as áreas i e j são vizinhas ou não. Valores do índice local de Moran significativamente diferentes do resto dos dados observados indicam áreas que possuem uma dinâmica espacial própria, ou seja, onde a suposição de estacionariedade pode não ser válida (Druck *et al.*, 2004).

2.2.2 Modelo de regressão espacial com efeito local

O modelo local adotado é a Regressão Geograficamente Ponderada (RGP), ou do inglês *Geographically Weighted Regression* (GWR). A idéia da *RGP* é ajustar uma regressão em cada ponto da região de estudo, ponderada pelos demais vizinhos por uma função de distância. A *RGP* é expressa por Fotheringham *et al.* (2002) como

$$\mathbf{Y}_{(s)} = \beta_{(s)}\mathbf{X} + \varepsilon, \tag{2.4}$$

onde $\mathbf{Y}_{(s)}$ é o fenômeno em estudo no ponto s e $\beta_{(s)}$ são os parâmetros a serem estimados no ponto s. A estimação de tais parâmetros via mínimos quadrados ponderados é dada por (Fotheringham *et al.*, 2002)

$$\hat{\beta}_{(s)} = \left(\mathbf{X}^T\mathbf{W}_{(s)}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}_{(s)}\mathbf{Y}, \tag{2.5}$$

onde $\mathbf{W}_{(s)}$ é um ajuste local que faz com que as observações mais próximas do ponto s tenham um peso maior. De fato, tome $\mathbf{W}_{(s)}$ tal que:

$$\mathbf{W}_{(s)} = \begin{bmatrix} w_{s1} & 0 & ... & 0 \\ 0 & w_{s2} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & w_{sn} \end{bmatrix} \tag{2.6}$$

Assim, nota-se que fazendo $w_{s1} = w_{s2} = ... = w_{sn} = 1$, *i.e.* dando o mesmo peso a todos os pontos, a matriz diagonal $\mathbf{W}_{(s)}$ fica igual à matriz identidade, e a estimação dos parâmetros fica igual à do modelo clássico de regressão linear.

O segredo então desse modelo de regressão espacial está em determinar os pesos w_{si} da matriz $\mathbf{W}_{(s)}$. Para tanto, faz-se uso de uma função de ponderação, ou $kernel$ em inglês. A Tabela 2.1 apresenta os principais modelos que estimam esses pesos.

Tabela 2.1: Funções de Ponderação

Modelo	**Função**
Gaussiano Fixo	$w_{si} = e^{-\frac{1}{2}\left(\frac{d(s,i)}{b}\right)^2}$
Gaussiano Variável	$w_{si} = e^{-\frac{1}{2}\left(\frac{d(s,i)}{b_{s(k)}}\right)^2}$
Biquadrático Fixo	$w_{si} = \begin{cases} \left[1-\left(\frac{d(s,i)}{b}\right)^2\right]^2 & , d(s,i) < b \\ 0 & , \text{caso contrário} \end{cases}$
Biquadrático Variável	$w_{si} = \begin{cases} \left[1-\left(\frac{d(s,i)}{b_{s(k)}}\right)^2\right]^2 & , d(s,i) < b_{s(k)} \\ 0 & , \text{caso contrário} \end{cases}$

Nota: o parâmetro $d(s,i)$ representa a distância entre o ponto s e o ponto i, e b é um parâmetro de suavização (ou $bandwidth$, em inglês) que assume duas formas: fixo e variável. O parâmetro de suavização fixo b representa uma distância fixa, enquanto que o parâmetro de suavização variável $b_{s(k)}$ representa uma distância condicionada a um número fixo de vizinhos k, $i.e.$, uma distância tal que existam k pontos cuja distância ao ponto s é menor do que $b_{s(k)}$. Assim, o parâmetro de suavização variável é definido por k.

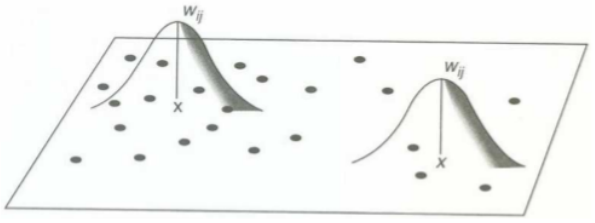


Figura 2.3: Parâmetro de suavização fixo.
Fonte:Fotheringham *et al.* (2002)

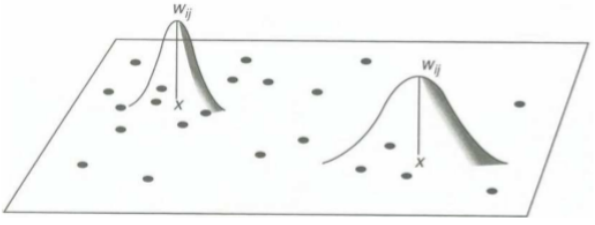


Figura 2.4: Parâmetro de suavização variável.
Fonte:Fotheringham *et al.* (2002)

Percebe-se pela Tabela 2.1 e pelas Figuras 2.3 e 2.4 que a determinação do parâmetro de suavização b influencia diretamente as variações locais dos parâmetros do modelo. Em outras palavras, um parâmetro de suavização muito grande diminui a variabilidade espacial dos parâmetros e faz com que o modelo local tenda a um modelo global, enquanto que um parâmetro de suavização menor pode captar melhor a estrutura espacial dos dados mas resulta em estimativas dos parâmetros com maiores erros padrão.

Assim, duas maneiras de determinar o parâmetro de suavização ótimo são dadas por (Fotheringham *et al.*, 2002)

$$\arg_b \, min \left\{ CV(b) = \sum_{i=1}^{n} \left(y_i - \widehat{y}_{\neq i}(b) \right)^2 \right\}. \tag{2.7}$$

Em geral, a validação cruzada, ou *Cross-Validation* em inglês, é utilizada quando a distribuição dos dados é a Gaussiana. Nesse método, a observação y_i não faz parte do ajuste do modelo no ponto i. Ou, de maneira geral, pode-se escolher o parâmetro de suavização tal que o critério de informação de Akaike corrigido (AIC_c em inglês) seja minímo, *i.e.*:

$$\arg_b \, min \left\{ AIC_c(b) = 2n \ln(\widehat{\sigma}) + n \ln(2\pi) + \frac{n\left(n + tr\left(\widehat{\beta}\right)\right)}{n - 2 - tr\left(\widehat{\beta}\right)} \right\}, \tag{2.8}$$

onde n é a quantidade de observações e $tr()$ é o traço da matriz $\widehat{\beta}$ cujas linhas são iguais à $\widehat{\beta}_s, s = 1, ..., n$.

Nota-se, pela Equação 2.4, que os parâmetros β_i do modelo podem variar no espaço e, pela Equação 2.5, β_i pode ser estimado em qualquer ponto do espaço, *i.e.* tanto em pontos observado quanto em pontos não observados. Por outro lado, y_i só pode ser estimado nos pontos observados.

2.2.3 Modelo de regressão logística geograficamente ponderada

Se a variável em estudo for do tipo binária, o uso da RGP deve ser adaptado, utiliza-se em seu lugar a Regressão Logística Geograficamente Ponderada (RLGP). Neste caso, um modelo de regressão logística fornece a probabilidade de ocorrência do fenômeno em estudo. Logo, enxerga-se a variável resposta do modelo como sendo:

$$Y \sim \text{Bernoulli}(1, \pi). \tag{2.9}$$

Assim, uma maneira de modelar $E(Y|x)$ é definida por Hosmer e Lemeshow (2000) como:

$$E(Y|x_i) = \pi(x_i) = \frac{e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_i}}{1 + e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_i}}, \tag{2.10}$$

onde $\pi(x)$ é a função de ligação logística. A adoção da função logística para adequar a variável resposta a um modelo de regressão permite que, com a transformação logit o modelo possa ser escrito equivalentemente como:

$$logit(\pi(x)) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_i, \tag{2.11}$$

fazendo com que o modelo logístico mantenha algumas propriedades importantes da regressão linear clássica. Nota-se que, diferentemente do caso da regressão linear clássica, os efeitos dos parâmetros β_j, $j = 1, ..., k$, na média da variável resposta não são mais interpretados diretamente, *i.e.*, interpreta-se agora as variações desses parâmetros na razão de chances, ou *odds ratio* em inglês (OR).

Conforme descrito por Hosmer e Lemeshow (2000), a razão de chances é simplesmente

$$OR = \frac{\frac{\pi(x_1=1)}{1-\pi(x_1=1)}}{\frac{\pi(x_1=0)}{1-\pi(x_1=0)}} = \exp(\beta_1). \tag{2.12}$$

Assim, para $\beta > 0$, tem-se que $OR > 1$. Isto é, um parâmetro positivo faz com que a razão de chances seja maior do que 1, fazendo com que a probabilidade do evento em estudo aumente para valores maiores de x. E para o caso de $\beta < 0$, tem-se que $0 < OR < 1$. Ou seja, a probabilidade do evento em estudo diminui para valores maiores de x.

Utiliza-se o método da máxima verossimilhança para estimar os parâmetros do modelo, *i.e.*, procura-se obter estimações dos parâmetros tal que a probabilidade de se observar o conjunto de dados obtidos seja máxima. Assim, de acordo com a Equação 2.9 a função de máxima verossimilhança é dada por:

$$\mathrm{l}(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i}. \tag{2.13}$$

Aplicando a transformação logaritmo, a função de máxima log-verossimilhança pode ser escrita como:

$$\mathrm{L}(\beta) = \ln\left[\mathrm{l}(\beta)\right] = \sum_{i=1}^{n} \left\{ y_i \ln\left[\pi(x_i)\right] + (1 - y_i) \ln\left[1 - \pi(x_i)\right] \right\} \tag{2.14}$$

$$= \sum_{j=0}^{k} \sum_{i=1}^{n} \left(y_i \, \beta_j \, x_{ij} - \ln\left(1 + e^{\beta_j x_{ij}}\right) \right). \tag{2.15}$$

Em seguida, derivando em relação a cada β e igualando a zero, obtém-se um sistema de equações cuja solução, *i.e.* as estimativas dos parâmetros, é encontrada por métodos numéricos iterativos, como o método de *Newton-Raphson*.

A RLGP é uma extensão do modelo de regressão logística. Portanto, o modelo fornece a probabilidade de ocorrência do fenômeno em estudo para um determinado

ponto do espaço s, que é dada por (Fotheringham *et al.*, 2002)

$$\pi(Y_s) = \frac{e^{\beta_{0(s)} + \sum_{j=1}^{k} \beta_{j(s)} x_{i(s)}}}{1 + e^{\beta_{0(s)} + \sum_{j=1}^{k} \beta_{j(s)} x_{i(s)}}}, \tag{2.16}$$

ou, equivalentemente, como:

$$ln\left(\frac{\pi(Y_s)}{1-\pi(Y_s)}\right) = \beta_{0(s)} + \sum_{j=1}^{k} \beta_{j(s)} x_{i(s)}, \tag{2.17}$$

onde $\pi(Y_s)$ é a probabilidade estimada de ocorrência do fenômeno no ponto observado s, $\beta_{j(s)}$, $j = 0,...,k$, são os parâmetros do modelo a serem estimados em cada ponto s. E assim como na *RGP*, ajusta-se um modelo para cada ponto do modelo incorporando uma matriz $\mathbf{W}_{(s)}$ (2.6) que pondera os pesos das observações vizinhas. A função de máxima verossimilhança responsável pela estimação de tais parâmetros é dada por (Fotheringham *et al.*, 2002)

$$\ln\left[l(\beta_s)\right] = \sum_{j=0}^{k} \left(\sum_{i=1}^{n} w_{is}\, y_i\, x_{ij}\right) \beta_{js} - \sum_{i=1}^{n} w_{is} \ln\left(1 + e^{\sum_{j=0}^{k} \beta_{js} x_{ij}}\right). \tag{2.18}$$

E, como no caso da regressão logística, não existe forma fechada para o cáculo dos parâmetros estimados após a diferenciação em relação aos β_s. Portanto faz-se uso de métodos numéricos iterativos, como o método dos Mínimos Quadrados Reponderados Iterativos (MQRI).

Assim como na regressão logística, pode-se testar a significância dos parâmetros da *RLGP*. Em outras palavras, deseja-se testar as hipóteses

$$\begin{cases} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{cases} \tag{2.19}$$

Uma maneira de se fazer inferência para as variáveis do modelo de regressão logística é dada por Hosmer e Lemeshow (2000), considerando

$$\mathrm{W} = \frac{\widehat{\beta}}{\widehat{\mathrm{SE}}(\widehat{\beta})} \sim \mathrm{N}(0,1), \tag{2.20}$$

onde W é a estatística do *teste de Wald* sob a hipótese de H_0 ser verdadeira, $\widehat{\beta}$ é a estimativa do parâmetro e $\widehat{\mathrm{SE}}(\widehat{\beta})$ é o desvio padrão associado à estimativa do parâmetro obtido a partir da diagonal da matriz inversa de informação de Fisher.

No entanto, sua aplicação para modelos locais precisa ser adaptada pois realiza-se múltiplos testes simultaneamente, *i.e.*, um para cada ponto s no espaço. O teste continua o mesmo, mas para saber se a estatística obtida é significativa dado um nível α, faz-se necessário comparar o valor do teste com um α adaptado dado por da Silva e Fotheringham (2016) como:

$$\alpha_{adaptado} = \frac{\alpha_{adotado}}{\frac{p_e}{p}}, \tag{2.21}$$

onde $\alpha_{adotado}$ é o nível de significância requerido do teste, p_e é a quantidade de parâmetros efetivos do modelo local, p é a quantidade de parâmetros no modelo global. Note que, $p_e \geq p$ e, quando adota-se um parâmetro de suavização muito grande, o modelo local converge para um modelo global fazendo com que $p_e = p$. Neste caso, nenhuma adaptação faz-se necessária e, portanto, testa-se a significância dos parâmetros usualmente.

2.3 Análise de superfície

Seja um fenômeno Z contínuo no espaço gerado por um processo estocástico $\{Z(u), u \in A\}$ sendo u um ponto qualquer da região de estudo A. A análise de superfície baseia-se em uma amostra contendo valores de Z em diferentes pontos da região de estudo, com o intuito de reconstruir a superfície completa dos valores de Z (Cressie, 1991). Em outras palavras, a análise de superficie estima o valor de Z para cada ponto

no espaço. Pode-se enxergar a análise de superficie como um processo pontual com atributo, pois os pontos no espaço estão atribuídos ao valor de um fenômeno.

2.3.1 Modelo determinístico local

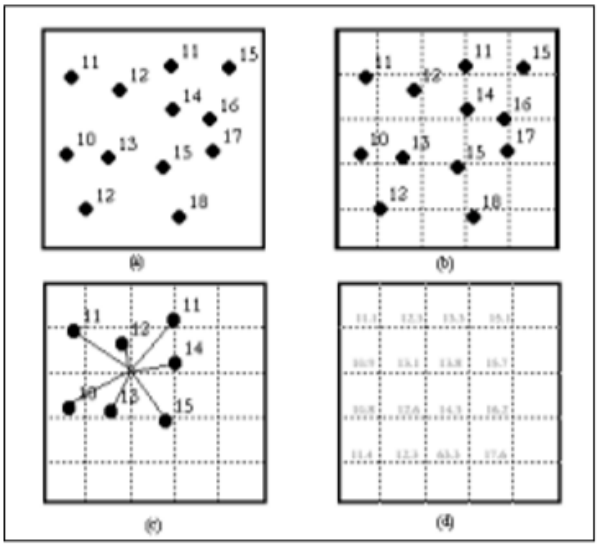


Figura 2.5: Interpolador local
Fonte:Druck *et al.* (2004)

Uma maneira simples de se interpolar uma superfície sem precisar parametrizar a estrutura de correlação-espacial dos dados, como é feito nos modelos de *krigagem*, é apresentado por Druck *et al.* (2004) como

$$\widehat{z_i} = \frac{\sum_{j=1}^{n} w_{ij} z_j}{\sum_{j=1}^{n} w_{ij}}, \tag{2.22}$$

onde $\widehat{z_i}$ é um ponto não observado sendo estimado pelos pontos vizinhos observados z_j, e w_{ij} é um fator de ponderação que faz com que observações mais próximas tenham pesos maiores. Adotou-se w_{ij} como sendo o inverso da distância ao quadrado entre z_i e z_j, mas existem diversas outras funções de ponderação para determinar os w_{ij}. Este modelo de estimação está representado na Figura 2.5.

Capítulo 3

MATERIAL E MÉTODOS

3.1 Introdução

Nesse Capítulo serão apresentados os materiais e métodos a serem utilizados no trabalho. Todas as bases que serão apresentadas e posteriormente utilizadas estão livres para *download* por qualquer pessoa. O tratamento e análise desses dados serão feitos utilizando os *softwares* SAS 9.4 e GWR4.

3.2 Material

A principal base de dados do trabalho é a de acidentes desagregados por ocorrência proveniente da Polícia Rodoviária Federal (PRF), responsável pelo sistema de registro de acidentes de trânsito das rodovias federais. Esta base de dados contém todos os acidentes nas rodovias federais no ano de 2017 registrados pela PRF. A Tabela 3.1 apresenta as variáveis da base de acidentes que foram utilizadas no modelo. Vale ressaltar que essas variáveis foram construídas a partir de condicionais das variáveis presentes na base de dados da PRF.

Tabela 3.1: Variáveis utilizadas no modelo

Variável	Valor	Descrição
feriu	1	Acidente com ao menos 1 ferido
	0	Caso contrário
chuva_garoa	1	Acidente sob chuva ou garoa
	0	Caso contrário
nevoeiro_neblina	1	Acidente sob nevoeiro ou neblina
	0	Caso contrário
noite	1	Acidente cuja fase do dia é em plena noite
	0	Caso contrário
mao_simples	1	Acidente cujo tipo de pista é simples
	0	Caso contrário
geometria_desfavoravel	1	Acidente cujo traçado da via é uma curva, um desvio temporário ou uma interseção de vias
	0	Caso contrário
fds	1	Acidente cujo dia da semana é sexta-feira, sábado ou domingo
	0	Caso contrário
ferias	1	Acidente cujo mês é janeiro, fevereiro, março, junho, julho ou dezembro
	0	Caso contrário

Vale ressaltar também que trabalhos similares a este não eram possíveis de se fazer em anos anteriores, uma vez que 2017 foi o primeiro ano de implementação do novo sistema de registros da PRF. E uma das novidades na base de dados montada pela PRF é a adoção do georreferenciamento dos acidentes por meio de coordenadas *lat-long*. Ou seja, cada acidente agora pode ser identificado como um ponto no espaço, e por consequência isto permitiu toda a análise deste trabalho.

Essa mudança feita pela PRF tem que ser reconhecida e estimulada para que mais bases de dados tenham observações georreferenciadas, pois isto eleva a qualidade da base de dados e amplia substancialmente as possibilidades metodológicas de análise.

Como já foi visto, o espaço é incorporado na modelagem. Assim, é possível visualizar os resultados obtidos, de forma mais clara e ampla, no espaço, *i.e.*, em um mapa

da região de estudo. Portanto faz-se bastante uso de *shapefile*, que é um tipo de base de dados geoespacial que contém a informação geométrica da unidade observacional (são exemplos de *shapefiles*: rios, montanhas, estradas, fronteiras de países, *etc*...). Utilizou-se o *shapefile* da malha rodoviaria brasileira (disponibilizado pelo Ministério dos Transportes) e dos municípios (disponível no site do IBGE), que podem ser vizualiados na Figura 3.1.

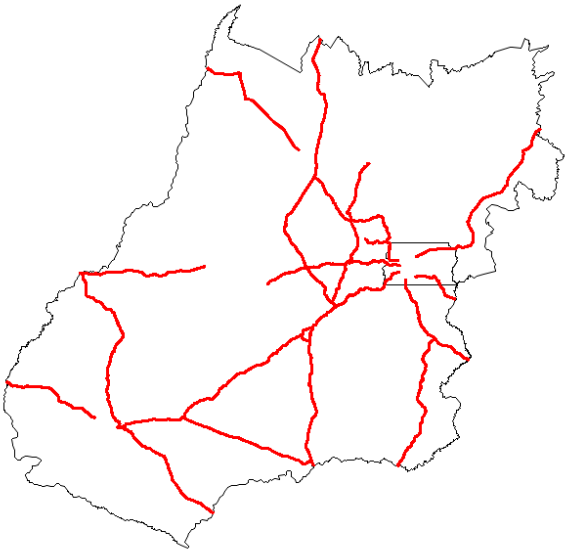


Figura 3.1: Rodovias Federais do Estado de Goiás e do DF

3.3 Métodos

3.3.1 Malha

Para poder analisar visualmente a distribuição espacial das estimativas dos parâmetros e da variável resposta, é preciso fazer essas estimativas para além dos pontos observados. No entanto, essas predições só fazem sentido ao longo das rodovias. Logo, o primeiro passo é criar uma área ao redor das rodovias e, dentro dessa área, gera-se uma quantidade n de pontos, com $n \to \infty$. Assim, obtém-se estimativas para "todo o

espaço" e consegue-se analisar os resultados obtidos de forma clara.

O passo a passo da construção dessa região está descrito na Figura 3.2, e pode ser descrito como:

- Passo 1: construir uma malha na região de estudo.
- Passo 2: selecionar os quadrados em que a linha está presente.
- Passo 3: selecionar os quadrados vizinhos aos selecionados anteriormente.

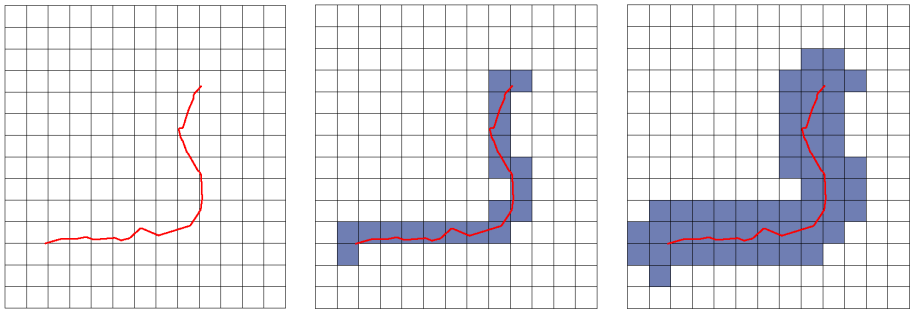


Figura 3.2: Passo a passo para a construção de uma área ao redor de uma linha.

3.3.2 Análise exploratória

O segundo passo é fazer a análise dos dados, que consiste primeiramente em uma análise exploratória para, em seguida, poder ajustar aos dados o melhor modelo de regressão.

Inicia-se a análise exploratória dos dados olhando para a distribuição e associação das variáveis por meio de tabelas e coeficientes de associação. Segundo Barbetta (2006), existem diversas medidas de associação entre variáveis que podem ser adotadas. Por exemplo, para variáveis quantitativas pode-se utilizar o coeficiente de correlação de Pearson, enquanto que para variáveis qualitativas nominais pode-se utilizar medidas baseadas no *Qui-Quadrado*, como o coeficiente de contingência modificado.

E, quando as variáveis são pelo menos qualitativas ordinais, pode-se utilizar a correlação de Spearman.

O coeficiente de contingência modificado é calculado (Barbetta, 2006) como

$$C = \sqrt{\frac{k\chi^2}{(k-1)(n+\chi^2)}} \tag{3.1}$$

onde n é a quantidade total de observações, k é o menor número de categorias entre as variáveis qualitativas e χ^2 é a estatística *Qui-Quadrado* dada por:

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}}, \tag{3.2}$$

onde i e j fazem referência à categoria das variáveis, O é a frequência observada e E é a frequência esperada. O coeficiente de contingência modificado varia entre 0 e 1, com valores próximos a 1 indicando que existe uma associação entre as variáveis.

Em seguida, verifica-se a depêndica espacial dos dados, conforme apresentado no capítulo anterior. E assim, pode-se definir um modelo para se ajustar aos dados.

3.3.3 Interpolação

O próximo passo é interpolar a probabilidade estimada pelo modelo para os pontos da região construída no primeiro passo. Os parâmetros estimados do modelo não precisam ser interpolados, pois os mesmos são estimados pelo modelo nesses mesmos pontos que compõem a região. Assim, os resultados obtidos podem ser apresentados por meio de mapas, ajudando a melhor entender como se relacionam as variáveis no espaço.

Capítulo 4

RESULTADOS

4.1 Introdução

Neste capítulo, são apresentados os resultados obtidos pelo uso da RLGP para modelar os acidentes de trânsito nas rodovias federais do estado de Goiás e DF. Tomou-se a variável *feriu* como a variável resposta e as demais variáveis da Tabela 3.1 como explicativas.

Para efeito de comparação, também é apresentado o ajuste do modelo de regressão logística clássico.

4.2 Análise exploratória

A Figura 4.1 traz a informação da distribuição espacial dos 5451 acidentes observados segundo a variável *feriu*. Percebe-se que, ao longo de quase todas as rodovias, houve algum tipo de acidente, e que na maioria deles tem pessoas feridas. De fato, pela Tabela 4.1, tem-se que em quase 65% dos acidentes existem feridos. No entanto, a partir dessas informações ainda não é possível observar algum padrão das ocorrências dos acidentes, e tampouco é possível identificar os trechos mais propensos a acidentes que resultam em feridos.

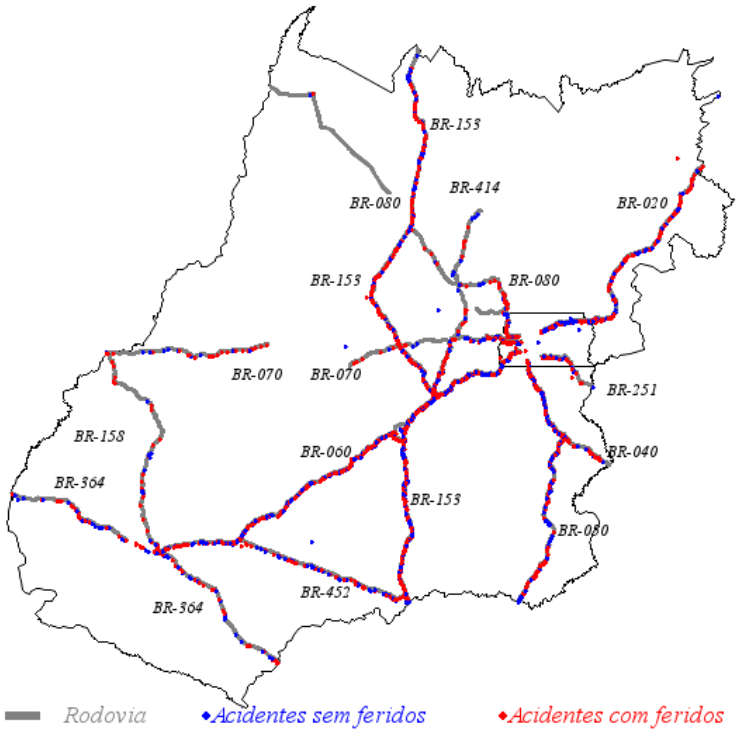


Figura 4.1: Acidentes registrados nas rodovias federais do DF e Goiás em 2017

Tabela 4.1: Tabela de frequência: *feriu*

feriu	**Frequência**	**Porcentagem**
0	1923	35,28
1	3528	64,72
Total	5451	100,00

Assim, procura-se verificar a associação entre a variável *feriu* com as demais variáveis apresentadas na Tabela 3.1. Em um primeiro momento, analisa-se o grau de dependência das variáveis pela Tabela 4.2. Percebe-se que de maneira geral, as variáveis possuem uma baixa associação entre elas. E o fato de que apenas 0,13 seja o maior coeficiente de contingência modificado entre a variável resposta *feriu* e as demais variáveis do modelo, indica que provavelmente o modelo não terá um bom ajuste. Em outras palavras, talvez as variáveis selecionadas não consigam explicar o fenômeno em estudo.

Tabela 4.2: Associação das variáveis segundo o coeficiente de contingência modificado

Variável	***feriu***	***chuva_ garoa***	***noite***	***mao_ simples***	***geometria_ desfavoravel***	***fds***	***ferias***	***nevoeiro_ neblina***
feriu	-							
chuva_garoa	0,13	-						
noite	0,05	0,05	-					
mao_simples	0,00	0,03	0,03	-				
geometria_desfavoravel	0,04	0,17	0,06	0,08	-			
fds	0,02	0,01	0,10	0,02	0,04	-		
ferias	0,03	0,17	0,09	0,00	0,02	0,05	-	
nevoeiro_neblina	0,03	0,04	0,02	0,03	0,02	0,01	0,02	-

Tabela 4.3: Tabela de contingência: *feriu* por *chuva_garoa*

feriu		*chuva_garoa*		
		0	1	Total
0	Frequência	1547	376	1923
	Porcentagem (total)	28,38	6,90	35,28
	Porcentagem (coluna)	33,44	45,58	
1	Frequência	3079	449	3528
	Porcentagem (total)	56,49	8,24	64,72
	Porcentagem (coluna)	66,56	54,42	
Total	Frequência	4626	825	5451
	Porcentagem	84,87	15,13	100,00

Analisando agora as distribuições das variáveis, pode-se verificar na Tabela 4.3 que aproximadamente 15% dos acidentes ocorreram sob chuva. Entre esses acidentes, percebe-se que 54,42% envolveram algum ferido. Porcentagem bem menor do que a observada entre os acidentes que não ocorreram sob chuva (66,56%). Esta informação é um indício, que provavelmente será captado pelo modelo, que *feriu* e *chuva_garoa* estão associadas.

Tabela 4.4: Tabela de contingência: *feriu* por *nevoeiro_neblina*

feriu		*nevoeiro_neblina*		
		0	1	Total
0	Frequência	1911	12	1923
	Porcentagem (total)	35,06	0,22	35,28
	Porcentagem (coluna)	35,22	48,00	
1	Frequência	3515	13	3528
	Porcentagem (total)	64,48	0,24	64,72
	Porcentagem (coluna)	64,78	52,00	
Total	Frequência	5426	25	5451
	Porcentagem	99,54	0,46	100,00

Analisando agora a Tabela 4.4, tem-se que *nevoeiro_neblina* possui uma distribuição semelhante à variável *chuva_garoa*. De fato, 52% dos acidentes que ocorreram sob nevoeiro ou neblina envolveram algum ferido. Enquanto que 64,78% é a porcentagem dos acidentes com feridos que não ocorreram sob os efeitos de *nevoeiro_neblina*. No entanto, observou-se apenas 25 acidentes sob nevoeiro ou neblina.

Tabela 4.5: Tabela de contingência: *feriu* por *noite*

feriu		*noite*		
		0	1	Total
0	Frequência	1275	648	1923
	Porcentagem (total)	23,39	11,89	35,28
	Porcentagem (coluna)	34,13	37,78	
1	Frequência	2461	1067	3528
	Porcentagem (total)	45,15	19,57	64,72
	Porcentagem (coluna)	65,87	62,22	
Total	Frequência	3736	1715	5451
	Porcentagem	68,54	31,46	100,00

Pela Tabela 4.5, percebe-se que 31,46% dos acidentes ocorreram em plena noite. A fase do dia parece não influenciar a ocorrência do tipo de acidente. Pois dado que o acidente ocorreu em plena noite, em 37,78% dos acidentes não houve feridos, enquanto que foi observado 34,13% de acidentes sem feridos quando a fase do dia não

era plena noite.

Tabela 4.6: Tabela de contingência: *feriu* por *mao_simples*

feriu		*mao_simples*		
		0	1	Total
0	Frequência	1366	557	1923
	Porcentagem (total)	25,06	10,22	35,28
	Porcentagem (coluna)	35,35	35,10	
1	Frequência	2498	1030	3528
	Porcentagem (total)	45,83	18,90	64,72
	Porcentagem (coluna)	64,65	64,90	
Total	Frequência	3864	1587	5451
	Porcentagem	70,89	29,11	100,00

Pela Tabela 4.6, observa-se que 29,11% dos acidentes ocorreram em vias de mão única. O condicionamento de *feriu* por *mao_simples* parece não influenciar a ocorrência do tipo de acidente. Resultados semelhantes a este foram observados relacionando *feriu* pelas variáveis *fds* (Tabela 4.8) e *ferias* (Tabela 4.9).

Tabela 4.7: Tabela de contingência: *feriu* por *geometria_desfavoravel*

feriu		*geometria_desfavoravel*		
		0	1	Total
0	Frequência	1540	383	1923
	Porcentagem (total)	28,25	7,03	35,28
	Porcentagem (coluna)	34,61	38,22	
1	Frequência	2909	619	3528
	Porcentagem (total)	53,37	11,36	64,72
	Porcentagem (coluna)	65,39	61,78	
Total	Frequência	4449	1002	5451
	Porcentagem	81,62	18,38	100,00

Analisando o efeito da variável *geometria_desfavoravel*, percebe-se pela Tabela 4.7 que 18,38% dos acidentes ocorreram sob seu efeito. Nota-se uma ligeira diferença na ocorrência do tipo de acidente segundo *geometria_desfavoravel*.

Tabela 4.8: Tabela de contingência: *feriu* por *fds*

feriu		*fds*		
		0	1	Total
0	Frequência	995	928	1923
	Porcentagem (total)	18,25	17,02	35,28
	Porcentagem (coluna)	35,97	34,56	
1	Frequência	1771	1757	3528
	Porcentagem (total)	32,49	32,23	64,72
	Porcentagem (coluna)	64,03	65,44	
Total	Frequência	2766	2685	5451
	Porcentagem	50,74	49,26	100,00

Tabela 4.9: Tabela de contingência: *feriu* por *ferias*

feriu		*ferias*		
		0	1	Total
0	Frequência	905	1018	1923
	Porcentagem (total)	16,60	18,68	35,28
	Porcentagem (coluna)	34,13	36,37	
1	Frequência	1747	1781	3528
	Porcentagem (total)	32,05	32,67	64,72
	Porcentagem (coluna)	65,87	63,63	
Total	Frequência	2652	2799	5451
	Porcentagem	48,65	51,35	100,00

Vale ressaltar que, como a unidade observacional é o acidente, *i.e.* um ponto no espaço, associado à informação de que houve ou não feridos envolvidos, não foi possível calcular o indicador de dependência espacial I de Moran. No entanto, é plausível supor que as variáveis não se comportem da mesma maneira no espaço e que haja uma certa estrutura de correlação espacial presente nos dados. A Figura 4.2 mostra diferentes trechos de rodovias presentes na região de estudo.

(a) BR-060

(b) BR-050

(c) BR-070

(d) BR-158

Figura 4.2: Diferentes trechos de rodovias federais no estado de Goiás e DF
Fonte: Google *Street view*

Verifica-se pela Figura 4.2 que existem rodovias duplicadas separadas por um canteiro central (Figura 4.2a), trechos em centros urbanos (Figura 4.2b), vias de mão simples bem sinalizadas (Figura 4.2d) ou rodovias que passam por reformas (Figura 4.2c). Estes exemplos representam uma pequena amostra do que pode ser encontrado ao longo das rodovias federais. Assim, é razoável supor que um acidente na BR-060 não é igual a um acidente na BR-158, por exemplo. Além do mais, estudos como o de Zheng *et al.* (2011) indicam que é mais pertinente modelar acidentes de trânsito por meio da estatística espacial local.

4.3 Modelo de regressão logística geograficamente ponderada

Ajustaram-se aos dados 5 modelos com a mesma estrutura: 1 modelo de Regressão Logística (RL) clássica e 4 modelos de Regressão Logística Geograficamente Ponderadas (RLGP), um para cada função de ponderação. Utilizou-se o algoritmo *Golden Section Search* para encontrar os valores dos parâmetros de suavização que minimizam a função (2.8). Algumas medidas de ajustes podem ser vistas na Tabela 4.10.

Tabela 4.10: Medidas de Ajustes dos Modelos Propostos

Modelo	**Função de ponderação**	**Parâmetro de suavização**	**Número de parâmetros**	**AIC_c**	**Variabilidade explicada**
RL	-	-	8	7033,08	0,008
RLGP	Gaussiana Fixa	140,09 km	21,47	7020,56	0,014
RLGP	Gaussiana Variável	1494 vizinhos	23,68	7007,01	0,017
RLGP	Biquadrática Fixa	387,68 km	17,64	7022,40	0,013
RLGP	Biquadrática Variável	2878 vizinhos	40,31	7013,77	0,020

Nota: no caso da RLGP, tem-se o número de parâmetros efetivos.

Selecionou-se os modelos RL e RLGP cuja função de ponderação é a *Gaussiana variável*, pois esta obteve o menor AIC_c. Pode-se perceber que todos os modelos apresentaram uma baixa variabilidade explicada, e que a RL se mostrou a menos eficiente.

A Tabela 4.11 mostra os resultados obtidos pelo modelo RL. Adotando um nível de significância de 10%, observa-se que apenas o *intercepto* e as variáveis *noite* e *chuva_garoa* possuem um efeito significativo no modelo. Percebe-se por esses parâ-

metros estimados que os acidentes que ocorrem em plena noite ou sob chuva e garoa possuem probabilidades menores de gerarem feridos.

De fato, tanto a baixa varibilidade explicada pelos modelos quanto os resultados obtidos podem ser explicados pelo fato de que, segundo PRF (2017), aproximadamente 40% dos acidentes possuem como causa uma falta de atenção dos envolvidos.

Assim, pode-se inferir pela razão de chances que acidentes em plena noite possuem uma chance 18% menor de gerarem feridos e acidentes sob chuva ou garoa possuem 39% a menos de chance de gerarem feridos pois nestas situações, que a princípio seriam de maior risco, os motoristas assumem uma postura mais defensiva ao volante.

Por fim, optou-se por não retirar as demais variáveis do modelo pois, desejou-se verificar a significância local das variáveis no modelo RLGP.

Tabela 4.11: Resultados do Modelo de Regressão Logística Clássica

Variável	**Valor estimado**	**Erro padrão**	**Wald z**	**p – valor**	**Razão de chances**
$Intercepto$	0,76	0,058	13,18	<,0001	-
fds	0,08	0,057	1,46	0,1458	1,09
$noite$	-0,19	0,061	-3,14	0,0017	0,82
$chuva\_garoa$	-0,50	0,078	-6,42	<,0001	0,61
$nevoeiro\_neblina$	-0,62	0,402	-1,54	0,1241	0,54
$mao\_simples$	0,01	0,063	0,20	0,8442	1,01
$geometria\_desfavoravel$	-0,11	0,074	-1,47	0,1413	0,90
$ferias$	-0,07	0,058	-1,14	0,2545	0,94

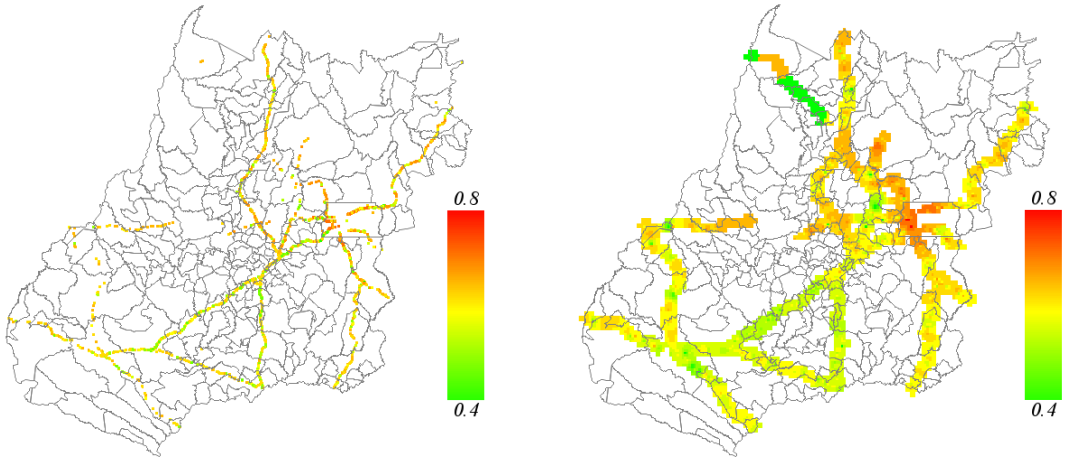
Os resultados obtidos do modelo RLGP podem ser vistos pela Tabela 4.12 e pelas Figuras 4.3, 4.4, 4.5 e 4.6. Começando pela Tabela 4.12, percebe-se que as estimativas de média/mediana dos parâmetros da RLGP são próximas das obtidas pela RL, *i.e.*, os parâmetros locais variam ao redor da estimativa global. É interessante notar que algumas variáveis possuem coeficientes que trocam de sinal no espaço, como é o caso

por exemplo de *noite*.

Tabela 4.12: Resultados do Modelo de Regressão Logística Geograficamente Ponderada

Variável	Mínimo	Primeiro quartil	Mediana	Média	Terceiro quartil	Máximo
Intercepto	0,610	0,677	0,821	0,879	1,105	1,234
fds	-0,007	0,014	0,038	0,042	0,075	0,090
noite	-0,372	-0,325	-0,198	-0,195	-0,099	0,011
chuva_garoa	-0,707	-0,648	-0,593	-0,595	-0,547	-0,468
nevoeiro_neblina	-1,251	-1,005	-0,559	-0,654	-0,397	-0,196
mao_simples	-0,173	-0,101	0,007	-0,008	0,077	0,146
geometria_desfavoravel	-0,416	-0,298	-0,100	-0,152	-0,024	0,037
ferias	-0,150	-0,107	-0,081	-0,087	-0,072	-0,033

As probabilidades estimadas de se ferir dada pela RLGP podem ser observadas na Figura 4.3. A Figura 4.3a apresenta as probabilidades estimadas pela RLGP nos pontos observados, enquanto que a Figura 4.3b traz a interpolação feita desses pontos para a região construída ao redor das rodovias. Percebe-se que tal procedimento facilita a análise dos resultados mantendo a essência do resultado original. Pode-se perceber também que as rodovias que ligam o Distrito Federal com o estado de Goiás são as mais propensas a gerarem feridos em um acidente.

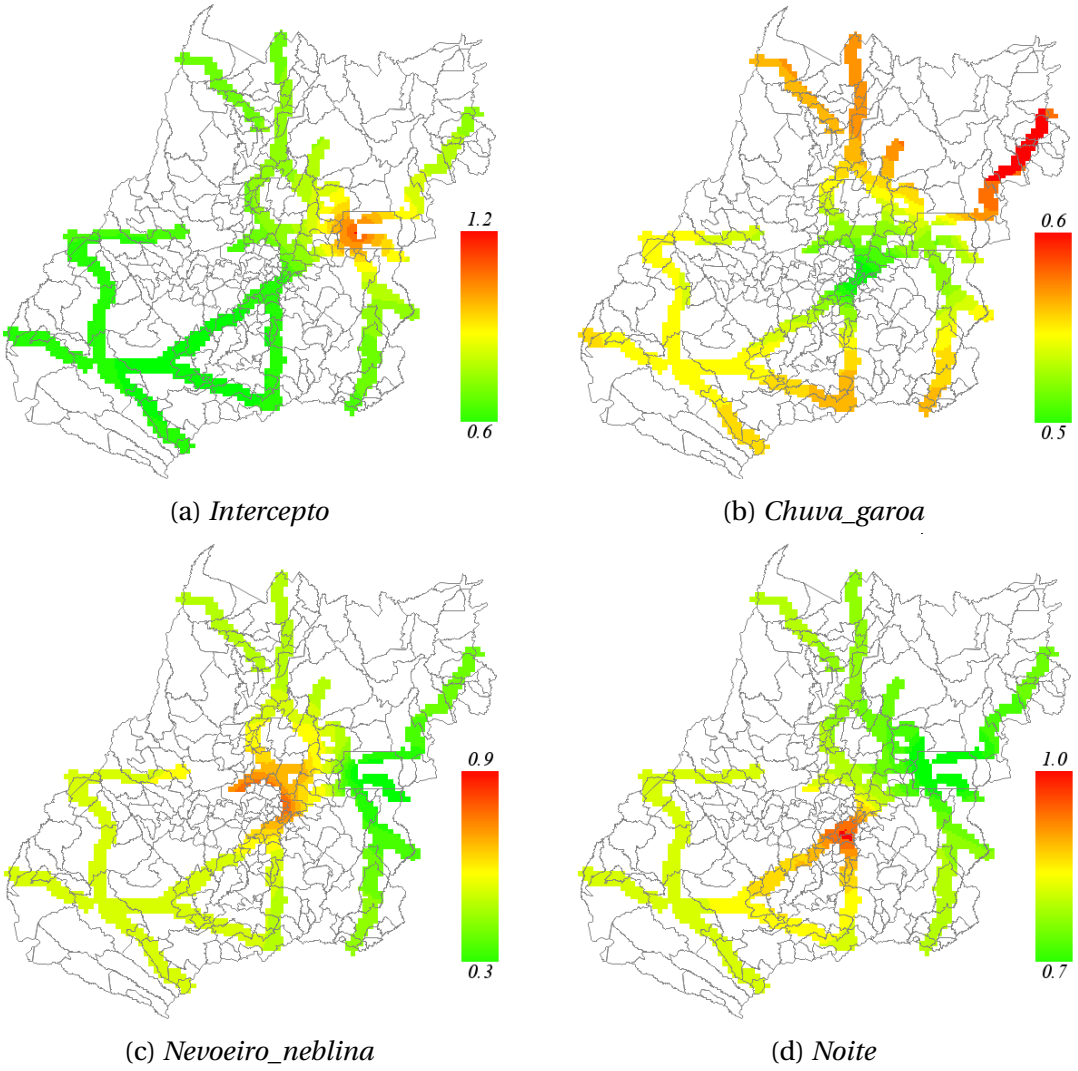


(a) Pontos observados (b) Interpolada

Figura 4.3: Probabilidades estimadas de se ferir em um acidente

A Figura 4.4 e a Figura 4.5 apresentam a distribuição espacial dos parâmetros. Para facilitar a interpretação dos resultados, decidiu-se por exibir a razão de chances ao invés da estimativa do parâmetro, exceto no caso do *intercepto*, em que é exibida a estimativa do parâmetro.

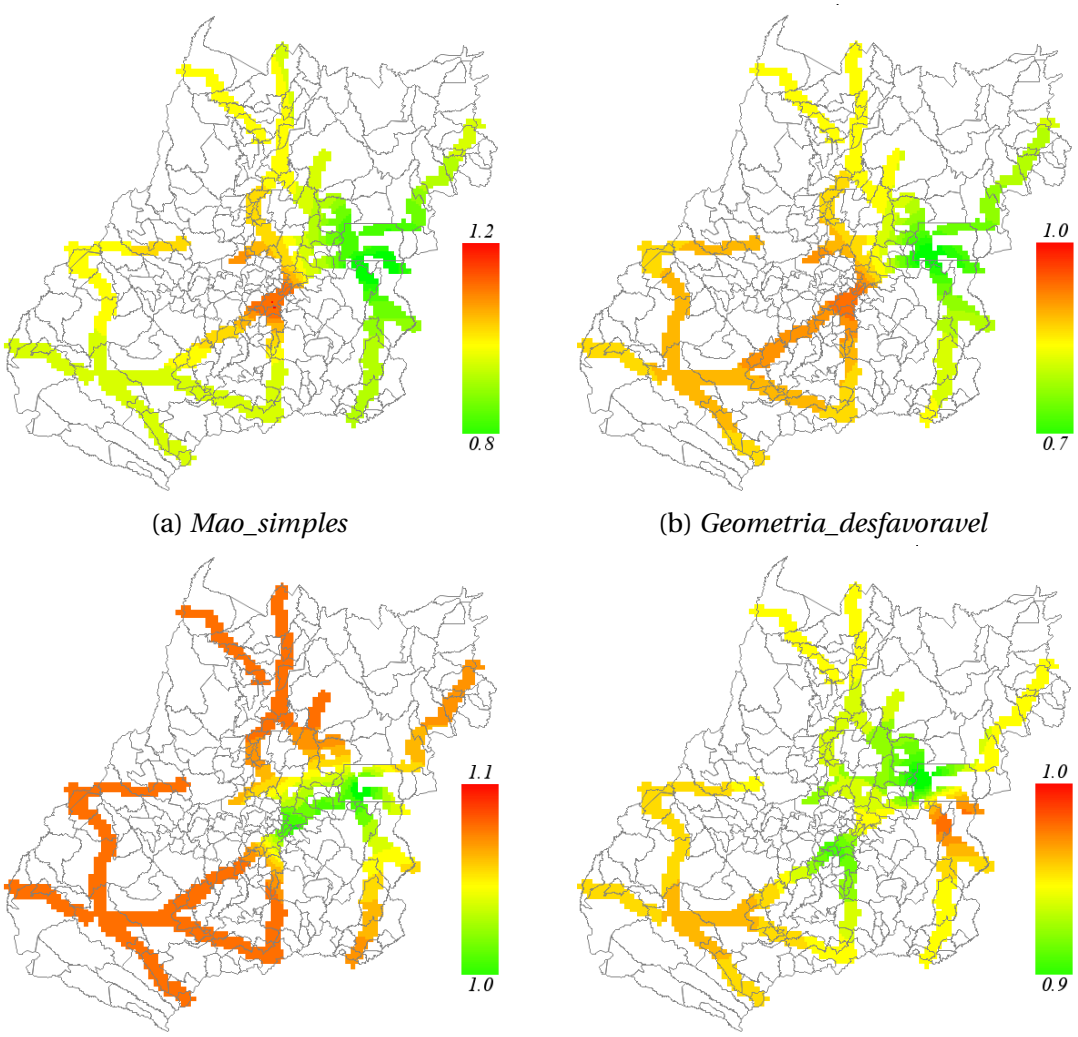
Percebe-se que cada variável tem sua própria dinâmica espacial, sendo mais ou menos intensa em determinada área da região em estudo.



(a) *Intercepto* (b) *Chuva_garoa*

(c) *Nevoeiro_neblina* (d) *Noite*

Figura 4.4: Distribuição espacial dos parâmetros



(a) *Mao_simples* (b) *Geometria_desfavoravel*

(c) *Fds* (d) *Ferias*

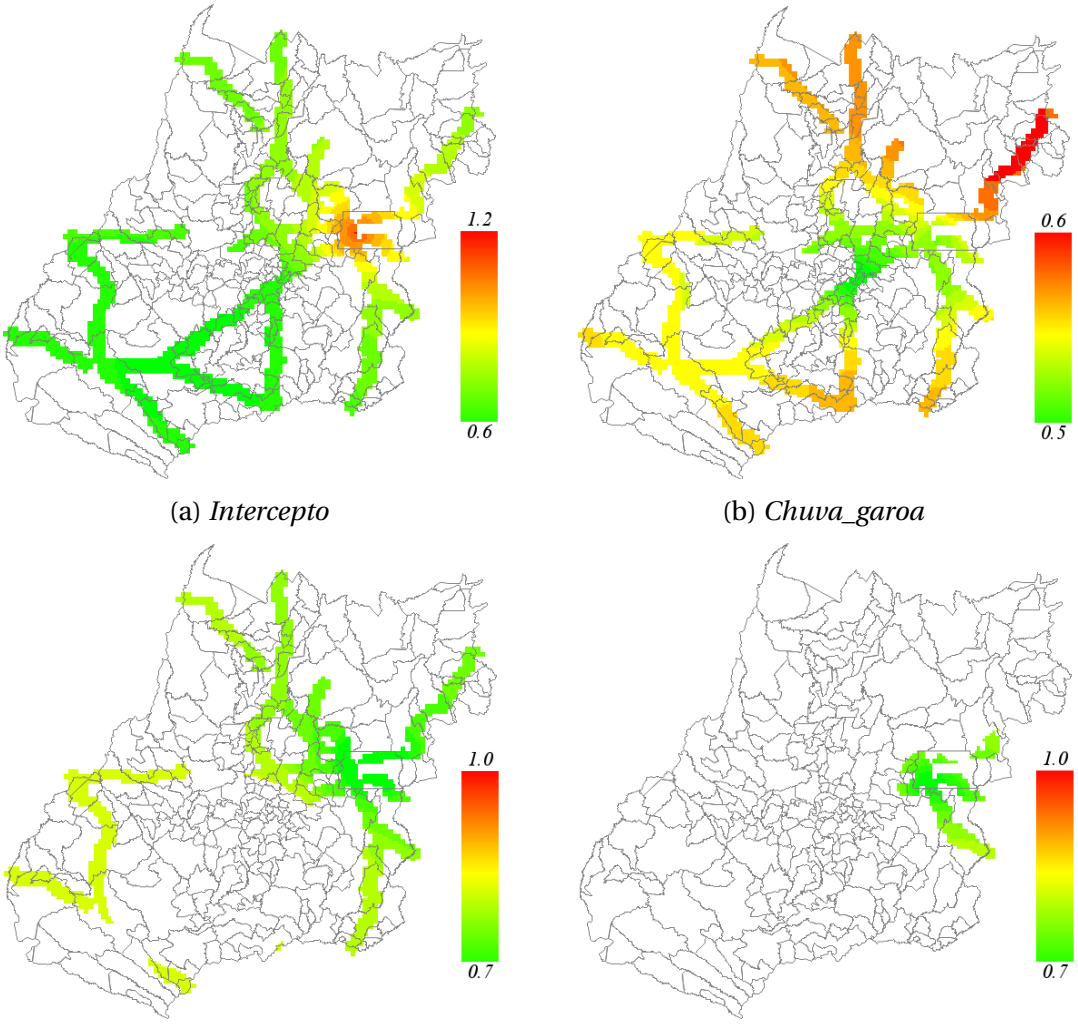
Figura 4.5: Distribuição espacial dos parâmetros

Tomando os mesmos 10% de significância adotados no modelo RL, resolveu-se testar a significância dos parâmetros estimados em todo o espaço. Lembrando que para garantir esse nível de significância é preciso fazer uma adaptação no mesmo. Assim, tem-se que

$$\alpha_{adaptado} = \frac{\alpha_{usual}}{\frac{p_e}{p}} = \frac{0,10}{\frac{23,678}{8}} = 0,0338 \tag{4.1}$$

ou seja, serão consideradas estatisticamente significativas 10% as estimativas que tiverem o p-valor menor do que 0,0338, e não menor do que 0,10.

A Figura 4.6 mostra as estimativas que foram significativas no modelo RLGP.



(a) *Intercepto* (b) *Chuva_garoa*

(c) *Noite* (d) *Geometria_desfavoravel*

Figura 4.6: Distribuição espacial dos parâmetros significativos

Nota-se que as mesmas variáveis que se mostraram significativas no modelo RL, *i.e.*, *intercepto*, *chuva_garoa* e *noite* também se mostraram significativas no modelo RLGP. É interessante notar que *noite* não é significativa para todo espaço, e que além dessas variáveis, *geometria_desfavoravel* também se mostrou significativa em uma determinada área da região de estudo.

Por fim, pode-se analisar a matriz de confusão dos modelos RL e RLGP pela Tabela

4.13. Percebe-se que ambos os modelos superestimaram as probabibilidades de haver feridos em um acidente e que a RL classificou o acidente de maneira equivocada um pouco mais do que a RLGP.

Tabela 4.13: Matriz de confusão

(a) RL

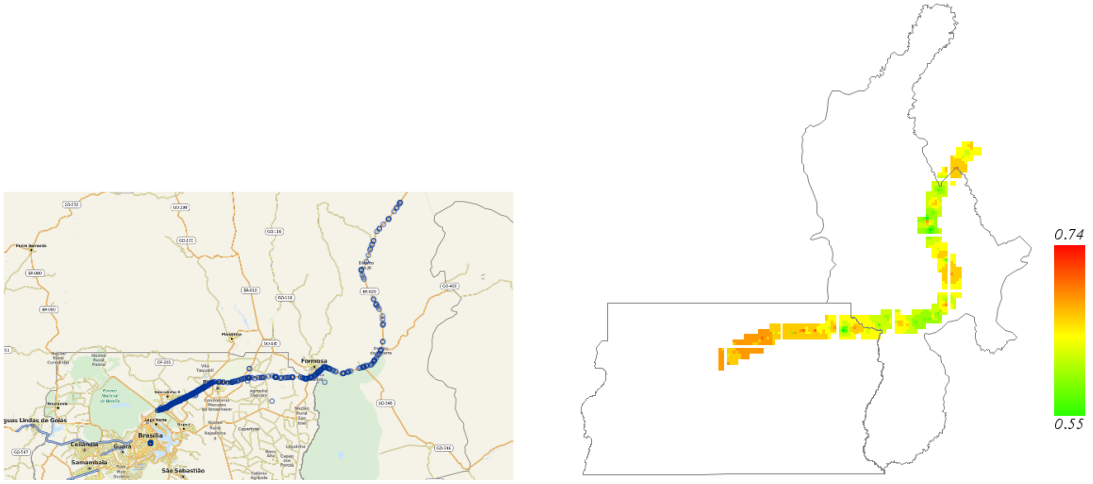
feriu	*feriu* estimado		
	0	1	Total
0	25	1898	1923
	0,46%	34,82%	35,28
1	25	3503	3528
	0,46%	64,26%	64,72
Total	50	5401	5451
	0,92	99,08	100,00

(b) RLGP

feriu	*feriu* estimado		
	0	1	Total
0	123	1800	1923
	2,26%	33,02%	35,28
1	98	3430	3528
	1,80%	62,92%	64,72
Total	221	5230	5451
	4,05	95,95	100,00

4.3.1 Análise de um trecho específico

Uma abordagem interessante de se fazer é selecionar apenas um trecho e olhar detalhadamente como se comportam as probabilidades de acidentes com feridos. Assim, selecionou-se um trecho da rodovia BR-030 que liga o DF com o estado de Goiás passando por Formosa.



(a) *Zoom da região no mapa* (b) *Zoom das probabilidades estimadas*

Figura 4.7: Estudo de caso

Esse trecho selecionado, que pode ser visto pela Figura 4.7, é particularmente interessante pois tem-se que 4 variáveis do modelo (*intercepto*, *chuva_garoa*, *noite* e *geometria_desfavoravel*) são estatisticamente significativas e que existe uma longa curva no trajeto. Nota-se que as probabilidades de acidentes com feridos nas condições observadas é maior no Distrito Federal, e em alguns pontos da curva que não eram possíveis de se observar na Figura 4.3.

Capítulo 5

CONCLUSÕES

Pode-se concluir que, tanto o modelo de regressão logística quanto o de regressão logística geograficamente ponderada adotados para se modelar a ocorrência de feridos em acidentes de trânsito nas rodovias federais do Distrito Federal e do estado de Goiás, se mostraram ineficientes. Vale notar que os dois modelos obtiveram resultados semelhantes, mas o modelo local levou vantagem sobre o global. Além disso, com o modelo RLGP foi possível analisar os resultados por meio de mapas e verificar a associação das variáveis no espaço.

Ao contrário do que se poderia supor inicialmente, obteve-se dos modelos que acidentes que ocorrem sob chuva, em plena noite ou em curvas possuem menos chances de gerarem feridos. Além disso, o fato de a pista ser simples não interfere na probabilidade de o acidente resultar em feridos. Em estudos futuros sobre acidentes em rodovias talvez seja pertinente incluir outros tipos de variáveis no modelo, como por exemplo, variáveis socioeconômicas. Uma outra alternativa para se modelar acidentes de trânsitos é agregar as observações (acidentes) por alguma unidade de área e adotar modelos locais de contagem, como, por exemplo, a regressão binomial negativa geograficamente ponderada ou a regressão de Poisson geograficamente ponderada.

Referências Bibliográficas

Anselin, L. (1996). The moran scatterplot as an esda tool to assess local instability in spatial association. *Spatial Analytical Perspectives on Gis in Environmental and Socio-Economic Sciences*.

Barbetta, P. A. (2006). *Estatística Aplicada às Ciências Sociais.*, (6 ed.). UFSC.

CNT (2017). Boletim estatístico - cnt - dezembro 2017. Technical report, Confederação Nacional dos Transportes. URL `http://www.cnt.org.br/Boletim/boletim-estatistico-cnt`. Acesso em 05 mar. 2018.

Cressie, N. A. C. (1991). *Statistics for Spatial Data*. Wiley.

da Silva, A. R. e Fotheringham, A. S. (2016). The multiple testing issue in geographically weighted regression. *Geographical Analysis*, 48:233–247.

DENATRAN (2017). Frota de veículos - 2017. Technical report, Departamento Nacional de Trânsito. URL `http://www.denatran.gov.br/index.php/estatistica/610-frota-2017`. Acesso em 08 mar. 2018.

Druck, S., Câmara, G., Monteiro, A. M. V., e Carvalho, M. S. (2004). *Análise Espacial de Dados Geográficos*. Embrapa.

Fotheringham, A. S., Brunsdon, C., e Charlton, M. (2002). *Geographically Weighted Regression: the analysis of spatially varying relationships*. Wiley.

Hosmer, D. W. e Lemeshow, S. (2000). *Applied logistic regression*. Wiley.

PRF (2017). Balanço - prf - 2017. Technical report, Policia Rodoviaria Federal. URL `https://www.prf.gov.br/portal/sala-de-imprensa/releases-1/balanco-prf-2017`.

SCHROEDER, E. M. e CASTRO, J. C. (1996). Transporte rodoviário de carga no brasil: situação atual e perspectivas. *Revista do BNDES*, 3:173–187.

Tobler, W. R. (1970). A computer model simulating urban growth in the detroit region. *Econ. Geo.*, 46:234–240.

Zheng, L., Robinson, R. M., Khattak, A., e Wang, X. (2011). All accidents are not equal: Using geographically weighted regressions models to assess and forecast accident impacts. *3 rd International Conference on Road Safety and Simulation*.