Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Dissertação de Mestrado

Análise da taxa de acidentes de trânsito com vítimas usando a Regressão Beta Geograficamente Ponderada

por

Roberto de Souza Marques Buffone

Brasília, Junho de 2023

Análise da taxa de acidentes de trânsito com vítimas usando a Regressão Beta Geograficamente Ponderada

por

Roberto de Souza Marques Buffone

Dissertação apresentada ao Departamento de Estatística da Universidade de Brasília, como requisito parcial para obtenção do título de Mestre em Estatística.

Orientador: Prof. Dr. Alan Ricardo da Silva

Brasília, Junho de 2023

Dissertação submetida ao Programa de Pós-Graduação em Estatística do Departamento de Estatística da Universidade de Brasília como parte dos requisitos para a obtenção do título de Mestre em Estatística.

Texto aprovado por:

Prof. Dr. Alan Ricardo da Silva

Orientador, EST/UnB

Profa. Dra. Terezinha Késsia de Assis Ribeiro

PGEST/UnB

Prof. Dr. Flávio José Craveiro Cunto

DET/UFC

Eu tenho fé, amor e afeto no século XXI, onde as conquistas científicas, espaciais, medicinais e a confraternização dos homens (...) serão as armas da vitória para a paz universal.

(Racionais MC's)

Para (e por) Deus.

Para (e por) minha família.

Para (e por) minhas mães Marias.

Mãe, pai e irmão, nós conseguimos mais uma vez!

Agradecimentos

Tantos são os agradecimentos que aqui não caberia destacá-los porém, com mais veemência agradeço:

A Deus e à intercessão de Maria, minha advogada e protetora;

A toda minha família, que me deu suporte nos momentos de maior necessidade. Em especial, minha mãe Maria do Carmo de Souza Marques que sempre me incentivou em todas as minhas escolhas e meu pai Giacomo Buffone, hoje não mais presente em corpo, mas certamente contente em espírito por minha conquista. Agradeço a meu irmão Marcone de Souza Marques, que sempre surgiu em minha vida como um segundo pai. Agradeço a minha dinda Bianca de Souza Marques, que me acolheu em sua casa no início da jornada, me tratando como um filho. Agradeço a meu primo e grande amigo Rodrigo de Souza Silva, que me deu as primeiras aulas de lógica de programação e a minha cunhada Érika Lopes de Carvalho de Souza Marques, que sempre me deu os melhores conselhos possíveis;

À minha amada companheira Ana Maria Andrade Barroso, principal motivadora durante todo o período como mestrando, sempre complacente com minha escolha;

À minha psicóloga Célia Conceição Fernandes Santos, que me motivou a me inscrever no processo seletivo de entrada no mestrado e me acompanhou durante todo o processo, nunca me permitindo desistir;

A meus amigos que sempre emanaram energias positivas, torcendo por minha conquista. A participação de vocês nas noites de bebedeiras foi fundamental para manter a sanidade mental;

Ao Clube de Regatas do Flamengo, que por ser muito grande, me deu emoções (felizes ou tristes) nos momentos de maior escuridão da minha jornada;

A Neco Osvaldo, Gabriela Flor e Brigite Eugênia, fugas nos momentos de maior desespero;

Às pessoas que moldaram a minha paixão pelas ciências exatas, sabendo que sem elas eu não chegaria aqui. Em especial, ao Professor Hilnei Macedo da Silva, professor de matemática do meu ensino médio. Saúdo a todo o corpo docente do Colégio Estadual Teotônio Marques Dourado Filho, da minha querida Morro do Chapéu;

À meu orientador, Professor Dr. Alan Ricardo da Silva, que me ajudou em todo o período acadêmico e me motivou a trilhar o caminho do "mais complicado". Tive com você minhas menores menções do mestrado, mas sem dúvidas, foram os cursos em que mais aprendi. Agradeço aos professores da banca avaliadora, Drª. Terezinha Késsia de Assis Ribeiro e Dr. Flávio José Craveiro Cunto, que compartilharam suas experiências para uma melhor execução deste trabalho;

À Universidade de Brasília, casa que me acolheu durante os 7 anos da minha jornada acadêmica;

Ao Sistema de Informações em Acidentes de Trânsito de Fortaleza (SIAT/FOR) que disponibilizou grande parte dos dados utilizados nesse trabalho;

À irrefutável ciência.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

A regressão linear clássica permite, de forma simples, que uma variável quantitativa contínua seja modelada a partir de outras variáveis. Porém, esse tipo de metodologia possui alguns pressupostos, como a independência entre as observações, que se ignorados trazem problemas metodológicos. Adicionalmente, nem todos os dados se adéquam à distribuição normal, necessitando assim de outros tipos de regressão para a modelagem. Com isso, a Regressão Beta Geograficamente Ponderada (RBGP) é apresentada com intuito de atribuir o fator da dependência espacial ao estudo, juntamente com a análise de taxas e proporções a partir da distribuição beta, que tem seu suporte no intervalo unitário e tem uma fácil adequabilidade, por seu ajuste flexível, aos dados estudados.

Neste trabalho a RBGP foi aplicada à taxa de acidentes de trânsito com vítimas em Fortaleza-CE, entre os anos de 2009 a 2011, comparando seus resultados aos modelos globais e locais de regressão clássica e de regressão clássica com a transformação da variável resposta pela função logito e à regressão beta global. Além disso, foi desenvolvido o pacote '*gwbr*' em R com os algoritmos necessários para a aplicação da RBGP.

Ao final, conclui-se que a abordagem local com o uso da distribuição beta é um modelo viável para explicar a taxa de acidentes de trânsito com vítimas, visto a adequabilidade do modelo tanto à distribuições assimétricas, quanto à distribuições simétricas. Por conta disso, se tratando da análise de taxas, é sempre recomendado o uso da distribuição beta.

Palavras-Chave: Acidentes de trânsito; Visão Zero; Dados espaciais; Regressão geograficamente ponderada; Regressão beta.

Abstract

Classical linear regression allows, in a simple way, that a continuous quantitative variable is modeled from other variables. However, this type of methodology has certain assumptions, such as independence between observations, which if ignored can lead to methodological issues. Additionally, not all data follows a normal distribution, which leads to alternative methods for modeling. In this context, Geographically Weighted Beta Regression (GWBR) is presented with the aim of incorporating spatial dependence into the modeling, along with the analysis of rates and proportions using the beta distribution. The beta distribution, with its scope within the unit interval and its flexible nature, easily adapts to the analyzed data.

In this study, GWBR was applied to the rate of traffic accidents with victims in Fortaleza-CE, Brazil, from 2009 to 2011, comparing its results to global and local models of classical regression, classical regression with logit transformation of the response variable, and global beta regression. Additionally, the '*gwbr*' package was developed in R software, providing the necessary algorithms for GWBR application.

In conclusion, it was found that the local approach using the beta distribution is a viable model for explaining the rate of traffic accidents with victims, given its suitability to both asymmetric and symmetric distributions. Therefore, when analyzing rates, the use of the beta distribution is always recommended.

Keywords: Traffic accidents; Zero Vision; Spatial data; Geographically weighted regression; Beta regression.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Em 1997, o parlamento sueco iniciou um debate sobre o programa Visão Zero que objetiva zero acidentes de trânsito graves e fatais (Johansson, 2009). Para atingir esse objetivo, o programa considera que diversos fatores contribuem para uma mobilidade segura, incluindo a geometria das estradas, as velocidades máximas definidas e a tecnologia aplicada ao estudo do tráfego e as políticas de controle (Vision Zero Network, 2014). Ainda assim os acidentes podem ocorrer, entretanto, os acidentes graves e fatais seriam menos observados. Diante disso, técnicas para estabelecer a relação entre duas ou mais variáveis em busca de uma modelagem de acidentes de trânsito são amplamente estudadas.

Dentre essas técnicas a regressão linear clássica é uma das metodologias estatísticas mais difundidas, permitindo que uma variável quantitativa contínua seja modelada a partir de outras variáveis. Esse método é muito utilizado por sua simplicidade e adequabilidade em diversas áreas tais como, pesquisa e mercado. Porém, a regressão linear clássica possui alguns pressupostos que por muitas vezes são ignorados e que conduzem a problemas metodológicos, gerando assim conclusões errôneas sobre o estudo. Suposições como a independência entre as observações e a distribuição gaussiana dos erros (ou da variável resposta), se não avaliadas, podem trazer resultados não acurados (Neter et al., 1983). Com isso, diversos outros métodos são estudados e desenvolvidos para se adequar a situações que não atendem aos pressupostos da

regressão linear clássica, tais como os modelos lineares generalizados e os modelos espaciais.

No caso de dados discretos, como por exemplo o número de acidentes de trânsito, a regressão linear clássica possui algumas limitações, como indicado por Chin e Quddus (2003); Miaou e Lum (1993); Jovanis e Chang (1986). Segundo Chin e Quddus (2003), o uso de uma regressão linear clássica para dados discretos pode incluir a presença de propriedades estatísticas indesejadas, como a possibilidade de uma contagem de acidentes negativa e a falta de ajuste à própria distribuição, por conta da assimetria comum aos dados supracitados. Nesse casos, é mais recomendado o uso dos modelos de regressão Poisson ou binomial negativo.

Quando o interesse está em modelar a taxa de acidentes **com vítimas**, um empecilho para o uso de um modelo de regressão para dados discretos é o valor ser contínuo e restrito ao intervalo [0,1]. Também não é adequado utilizar a regressão linear clássica, apesar do dado ser contínuo, porque os dados podem apresentar uma assimetria à direita, uma vez que a taxa de ocorrências com vítimas em geral é baixa em relação à quantidade de acidentes. Para situações como esta foi desenvolvido por Ferrari e Cribari-Neto (2004) o modelo de regressão beta, que considera que a variável resposta segue uma distribuição beta. Essa distribuição possui seu suporte definido no intervalo contínuo unitário (0,1), além de possuir flexibilidade para modelar dados simétricos e assimétricos.

A fim de incorporar o fator espacial ao estudo de incidentes de trânsito, Gomes et al. (2017) indicam o uso de modelos que consideram a dependência espacial, visto a influência entre eventos que ocorrem mais próximos uns dos outros. Essa estrutura de dependência também foi verificada por Obelheiro et al. (2020), Zhao e Park (2004), dentre outros.

Unindo os conceitos da regressão beta e regressão geograficamente ponderada, definido por Fotheringham et al. (2002), Da Silva e Lima (2017) desenvolveram a Regressão Beta Geograficamente Ponderada (RBGP) ou, em inglês, *Geographically Weighted Beta Regression* (GWBR). Essa abordagem busca modelar taxas ou proporções num contexto espacial.

Dessa forma, esse trabalho tem por objetivo a aplicação do modelo RBGP desenvolvido por Da Silva e Lima (2017) em acidentes de trânsito com ocorridos na cidade de Fortaleza, no Ceará,

entre os anos de 2009 e 2011, considerando a taxa de ocorrências com vítimas. Além disso, pretende-se criar um pacote no *software* R a partir da macro SAS desenvolvida por Da Silva e Lima (2017), o que permitirá uma maior visibilidade à técnica, visto a ampla utilização do *software* R.

O restante deste texto está organizado como se segue. No Capítulo 2 será desenvolvida uma revisão bibliográfica que embasa a motivação desse estudo, discutindo sobre acidentes de trânsito e o programa Visão Zero. No Capítulo 3 será apresentada uma revisão bibliográfica sobre o modelo de regressão beta geograficamente ponderado. O Capítulo 4 apresentará os materiais e métodos a serem utilizados no trabalho para ilustrar o uso do RBGP, e o Capítulo 5 apresenta a análise dos resultados. Por fim, no Capítulo 6 dar-se-ão as conclusões, limitações do trabalho e recomendações para trabalhos futuros.

Capítulo 2

Acidentes de Trânsito

2.1 Introdução

A segurança viária continua não sendo bem integrada aos planos estratégicos de tráfego, mesmo sendo uma das dimensões principais dos sistemas de transporte (Gomes et al., 2017). Segundo estudos apresentados por Chatterjee et al. (2001), Tarko (2006) e De Leur e Sayed (2002) isso ocorre principalmente devido a falta de dados e a limitação ferramental para as análises, resultando na necessidade de confiar em avaliações subjetivas de problemas de segurança viária.

Neste Capítulo serão descritos alguns pontos da abordagem tradicional da análise de acidentes de trânsito *versus* a Visão Zero. Além disso, serão descritos alguns fatores e técnicas atualmente utilizados para a modelagem de acidentes de trânsitos, tanto fatais quanto não-fatais.

2.2 Modelagem de acidentes

Quando revisada a literatura sobre as abordagens tradicionais para o planejamento de estradas e rodovias, mostra-se evidente a falta de consideração explícita de questões e preocupações de segurança no trânsito (De Leur e Sayed, 2002), como observado no processo de planejamento de transporte apresentado na Figura 2.1.

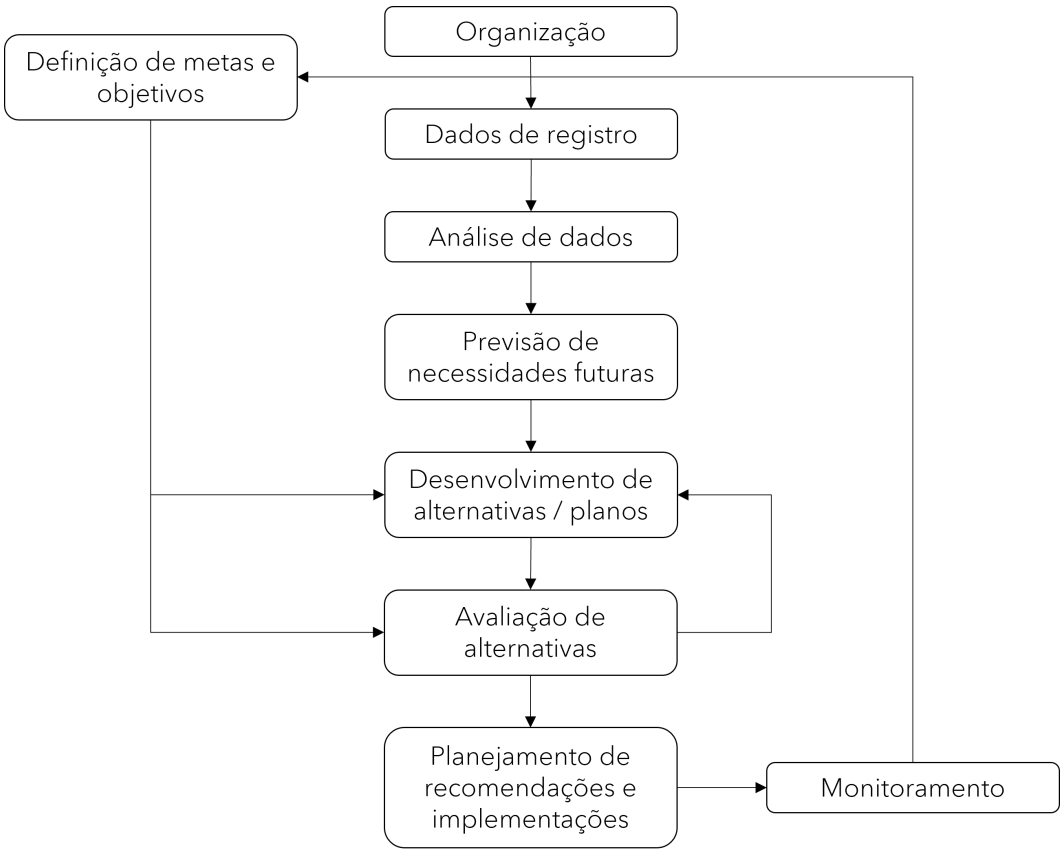

Figura 2.1: Principais etapas em um processo de planejamento de transporte.
Fonte: ITE (1982), com adaptações.

Em detrimento a isso, foi apresentado por Tarko (2006) um esquema que torna mais evidente a ideia de segurança no sistema de tráfego, como pode ser visto na Figura 2.2. Nesse novo esquema apresentado, a segurança é parte fundamental na construção da rede de transporte, sendo levada em consideração a cada etapa, seja na adição de novas informações de acidentes, ou até mesmo no recebimento de novos volumes de tráfego para a malha. Além disso, dentro do planejamento, também são avaliadas ações para uma segurança futura, tendo assim uma visão proativa para esse âmbito.

Uma vez evoluída a visão do planejamento de trânsito, são necessárias ferramentas que sirvam como sustentáculo para esse avanço, estimulando a busca por técnicas mais avançadas para que a modelagem de acidentes de trânsito seja feita de forma mais objetiva e precisa,

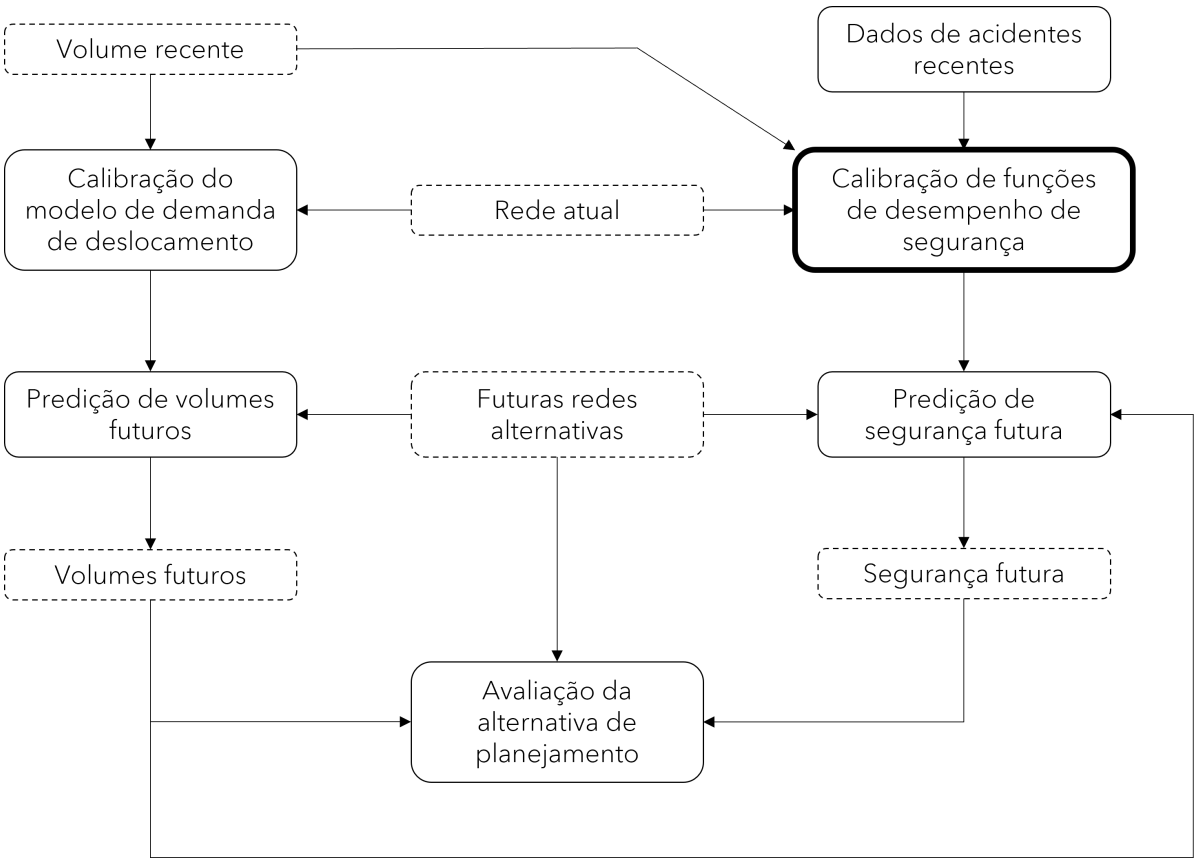

Figura 2.2: Previsão de segurança como parte da modelagem da rede de transporte.
Fonte: Tarko (2006), com adaptações.

podendo assim generalizar problemas ocorridos em busca de evitar tais incidentes. Para isso, vários são os estudos que visam modelar tais ocorrências, com diversas abordagens diferentes, como em Abdel-Aty e Radwan (2000); Quddus (2008); Xu e H. (2015); Gomes et al. (2017); Figueira et al. (2017).

Algumas características a serem consideradas durante a modelagem de acidentes de trânsito, buscando agir de forma proativa em relação a tais eventos, são a exposição ao risco (volume de tráfego, quilometragem), a probabilidade de envolvimento em um acidente de acordo com características pré-definidas e a severidade do acidente (De Leur e Sayed, 2002). Esse último tem extrema importância para o trabalho aqui desenvolvido, uma vez que é utilizada como ideia principal a estratégia Visão Zero que visa erradicar os acidentes graves e fatais (Johansson,

2009). Essa estratégia será vista mais a fundo na seção subsequente.

Para realizar a modelagem, deve-se considerar um agregamento espacial das ocorrências por unidades de área, como setor censitário, bairros ou, o mais comum em modelagens de acidentes de trânsito, as Zonas de Análise de Tráfego (ZAT), visto por exemplo em Quddus (2008); Siddiqui et al. (2012); Lee et al. (2014); Rhee et al. (2016); Gomes et al. (2017); Obelheiro et al. (2020).

As Zonas de Análise de Tráfego (ZAT) ou, em inglês, *'Traffic Analysis Zone'* (TAZ), ou apenas Zonas de Tráfego (ZT), são unidades geográficas construídas com base em *clusters*, de acordo com as características sociodemográficas da localidade (Martínez et al., 2009). O primeiro algoritmo sistemático voltado a definição de ZT's foi proposto por Openshaw (1977), otimizando uma função objetivo para partições de uma localidade de acordo com algumas variáveis observadas. Desde então, essa separação vem sendo uma das mais utilizadas para planos de transporte.

A frequência de acidentes pode então ser estimada para cada ZT de acordo com atributos associados como:

- Características da via: Volume de cruzamentos (Huang et al., 2010), vias com diferentes limites de velocidade (Abdel-Aty et al., 2011; Siddiqui et al., 2012), vias com diferentes classificações (Quddus, 2008; Hadayeghi et al., 2010; Huang et al., 2010), cruzamentos e rotatórias (Quddus, 2008);

- Padrão de tráfego em termos de volume e velocidade da via (Quddus, 2008; Hadayeghi et al., 2010);

- Origem e distribuição da rota (Abdel-Aty et al., 2011);

- Condições climáticas (Aguero-Valverde e Jovanis, 2006);

- Uso do solo (Hadayeghi et al., 2010; Siddiqui et al., 2012);

- Fatores socioeconômicos: Densidade populacional (Hadayeghi et al., 2006; Huang et al., 2010), idade (Aguero-Valverde e Jovanis, 2006; Quddus, 2008; Hadayeghi et al., 2010), renda familiar (Huang et al., 2010; Siddiqui et al., 2012; Xu et al., 2014) e emprego (Quddus, 2008; Hadayeghi et al., 2010; Huang et al., 2010).

Algumas propostas foram feitas para a modelagem de acidentes de trânsito, em que omite-se o fator espacial, como por exemplo o modelo linear generalizado com a distribuição binomial negativa (Hadayeghi et al., 2006; Aguero-Valverde e Jovanis, 2006; Quddus, 2008; Abdel-Aty et al., 2011) e o modelo Poisson lognormal Bayesiano (Siddiqui et al., 2012; Xu et al., 2014). Para modelos que consideram a dependência espacial, a literatura contempla uma abordagem Bayesiana (Aguero-Valverde e Jovanis, 2006; Quddus, 2008; Huang et al., 2010) e também modelos frequentistas, como modelos espaciais econométricos[1] (Quddus, 2008), a Regressão Poisson Geograficamente Ponderada (Hadayeghi et al., 2010) e a Regressão Binomial Negativa Geograficamente Ponderada (Gomes et al., 2017).

É visto que os fatores para os modelos supracitados incluem características do condutor do veículo e da via, confirmando então a necessidade de uma abordagem conjunta desses fatores para a construção do modelo de tráfego mais seguro, como indicado por Tarko (2006) e Chatterjee et al. (2001). Diante disso, diversas abordagens são tomadas e uma delas é a Visão Zero.

2.3 Visão Zero

Após as eleições suecas de 1994, o novo Ministro dos Transportes do país definiu como uma de suas prioridades a segurança viária. Foi então iniciado um diálogo entre a equipe do Ministério e o órgão de administração de estradas suecas (SRA) sobre como poderia se fazer da segurança no trânsito um assunto prioritário (Johansson, 2009).

Perante discussões e evoluções, no ano de 1997 o Parlamento Sueco aprovou um projeto

[1] Modelo Espacial Autoregressivo e Modelo de Erro Espacial (Anselin, 1988).

de lei sobre Segurança no Trânsito, definindo assim o Visão Zero: "Visão Zero significa que eventualmente ninguém será morto ou gravemente ferido no sistema de transporte rodoviário".

Essa abordagem, que requer uma mudança na forma como comunidades abordam decisões, ações e atitudes em torno da mobilidade segura, vem sendo adotada por um número crescente de comunidades ao redor do mundo e visa fundamentalmente a mudança de uma abordagem tradicional para uma pensamento mais moderno (Vision Zero Network, 2018), como pode ser observado na Tabela 2.1.

Tabela 2.1: Comparativo entre abordagem tradicional e Visão Zero.

Itens	**Abordagem**	
	Tradicional	**Visão Zero**
- Filosofia	- Acidentes são inevitáveis; - A mobilidade contém um percentual inevitável de ferimentos pessoais.	- Ninguém será morto ou gravemente ferido no sistema de transporte rodoviário; - As pessoas cometem erros, enganos e julgamentos errôneos; - Existem limites de tolerância biomecânica; - A cadeia de eventos pode ser cortada em muitos lugares.
- Moral indispensável	Não está claro.	Nunca é eticamente aceitável que pessoas sejam mortas ou feridas gravemente no sistema de transporte rodoviário.
- Problema	Tentar prevenir todas as lesões causadas pelo tráfego.	Evitar que acidentes resultem em vítimas graves e fatais.
- Meta apropriada	Prevenir acidentes no tráfego.	Eliminar fatalidades e ferimentos graves.
- Planejamento de abordagem	- Reativo a incidentes; - Abordagem incremental para reduzir o problema.	- Planejamento proativo; - Abordagem sistemática para construir um sistema rodoviário seguro; - Planejamento estratégico; - Planejamento operacional; - Planejamento tático.
- Causas do problema	Erro humano.	O design do sistema viário como causa principal e os projetistas do sistema como responsáveis.
- Foco no fator humano	Força mecânica excessiva para humanos.	Reduzir as forças mecânicas às tolerâncias humanas.
- Responsável final	Indivíduo usuário da via.	Responsabilidade dividida entre todos, incluindo aqueles que projetam, constroem, operam e utilizam o sistema viário.
- Metodologia de trabalho	Composta por intervenções isoladas.	Às vezes as pessoas cometem erros portanto, o sistema viário e as políticas relacionadas devem ser projetados para garantir que esses erros inevitáveis não resultem em ferimentos graves ou mortes.
- Custo de salvar vidas	Caro.	Barato.

Fonte: Safarpour et al. (2020), com adaptações.

A abordagem Visão Zero abre mão da ideia de erradicar com todo e qualquer tipo de acidente, focando em acidentes graves e fatais. Para isso, são presumidas responsabilidades conjuntas e em sequência, da seguinte forma:

1. Os engenheiros que projetam o sistema viário são responsáveis pelo nível de segurança por toda a malha;

2. Os usuários da via são responsáveis por seguir as regras de uso do sistema de transporte rodoviário definidas pelos projetistas do sistema;

3. Caso os usuários da estrada não cumprirem estas regras por falta de conhecimento, aceitação ou capacidade, ou se ocorrerem ferimentos, os projetistas do sistema são obrigados a tomar as medidas adicionais necessárias para neutralizar as pessoas mortas e gravemente feridas.

Uma das regras imputadas pelo Visão Zero, citada por Belin et al. (1997), que auxilia na definição de diretrizes para o trânsito, é o estudo do nível de violência que o corpo humano pode suportar sem ser morto ou gravemente ferido, tornando isso como parâmetro básico no projeto do sistema de transporte rodoviário.

É baseado nesse princípio que se pode desenvolver a futura sociedade com tráfego rodoviário seguro: através da concepção e construção de estradas e serviços de transporte de forma a não ultrapassar o nível de violência tolerável pelo ser humano; e através da contribuição efetiva de diferentes sistemas de apoio, como regras e regulamentos, educação, informação, vigilância, serviços de resgate, cuidados e reabilitação. Tendo isso como base, haverá uma demanda positiva por soluções novas e eficazes que possam contribuir para um sistema de transporte rodoviário onde as necessidades humanas estejam em foco (Johansson, 2009).

Segundo Vision Zero Network (2018), os elementos centrais definidos para o programa, que servem como principal base para o desenvolvimento do programa em outros países, são divididos em três áreas, da forma:

I. **Liderança e comprometimento**

1. **Compromisso público, em alto nível e contínuo:** Os gestores devem se comprometer com o objetivo de erradicar as mortes no sistema viário e definir metas objetivas para que isso ocorra.
2. **Compromisso autêntico:** Envolvimento da comunidade de maneiras significativas e culturalmente relevantes com o apoio de líderes comunitários.
3. **Planejamento estratégico:** Um Plano de ação Visão Zero é desenvolvido, aprovado e usado para orientar o trabalho. O Plano inclui metas explícitas e estratégias mensuráveis com cronogramas claros, identificando as partes interessadas responsáveis.
4. **Entrega de projeto:** Tomadores de decisão e engenheiros de tráfego promovem projetos e políticas, garantindo financiamento e implementando tais projetos, priorizando estradas com os problemas de segurança mais urgentes.

II. **Vias e velocidades seguras**

5. **Vias para todos:** Planos comunitários que são implementados por meio de projetos para incentivar uma rede de transporte segura e bem conectada para pessoas que usam todos os meios de transporte. Isso prioriza viagens seguras de pessoas em detrimento de viagens rápidas de veículos motorizados.
6. **Velocidades apropriadas ao contexto:** As velocidades de deslocamento são definidas e gerenciadas para alcançar condições seguras para o contexto específico da via e para proteger todos os usuários da via, principalmente aqueles com maior risco de acidentes.

III. **Abordagem baseada em dados, transparência e responsabilidade**

7. **Análise e programas com foco em ações:** O compromisso é feito com uma abordagem e resultados igualitários, incluindo a priorização de investimentos em comu-

nidades tradicionalmente mal atendidas.

8. **Planejamento proativo e sistêmico:** Identificar e abordar os principais fatores de risco e mitigar possíveis acidentes e sua gravidade.
9. **Planejamento responsivo e baseado em pontos frequentes:** Um mapa com os locais de acidentes graves e fatais na comunidade é desenvolvido, atualizado regularmente e usado para orientar ações prioritárias.
10. **Avaliação e ajustes compreensíveis:** A avaliação de rotina do desempenho das intervenções de segurança é tornada pública e compartilhada com os tomadores de decisão para informar prioridades, realocação de orçamentos e atualizações do Plano de Ação Visão Zero.

2.4 Modelagem de taxas de acidentes fatais

Segundo dados do Institute for Health Metrics and Evaluation (2019), em 2019 as mortes por acidentes nas estradas ocuparam a 12ª posição dentre as causas de mortes totais no mundo (Figura 2.3), ficando a frente por exemplo, de causas como a tuberculose (13º), o vírus HIV (14º) e os homicídios (17º).

Quando estudada apenas a população mais jovem, entre 5 e 49 anos, ainda para as ocorrências de 2019, existe uma mudança substancial nesse ranking, fazendo com que agora as mortes ocorridas em acidentes de trânsito fiquem em 3º lugar (Figura 2.4), atrás apenas de doenças cardiovasculares e neoplasias, como o câncer.

Observando as ocorrências para alguns países nos últimos anos e utilizando dados de World Health Organization (2019), tem-se os dados apresentados na Figura 2.5, que relativiza as ocorrências de acordo com o tamanho populacional, considerando o número de mortes a cada 100 mil habitantes. Um país exemplificado é a Tailândia, que mostra um crescimento considerável na taxa, chegando a atingir quase 30 mortes a cada 100 mil habitantes em 2019. Um outro país que possuía taxas altíssimas mas que vem conseguindo decair com esse valor é o Brasil,

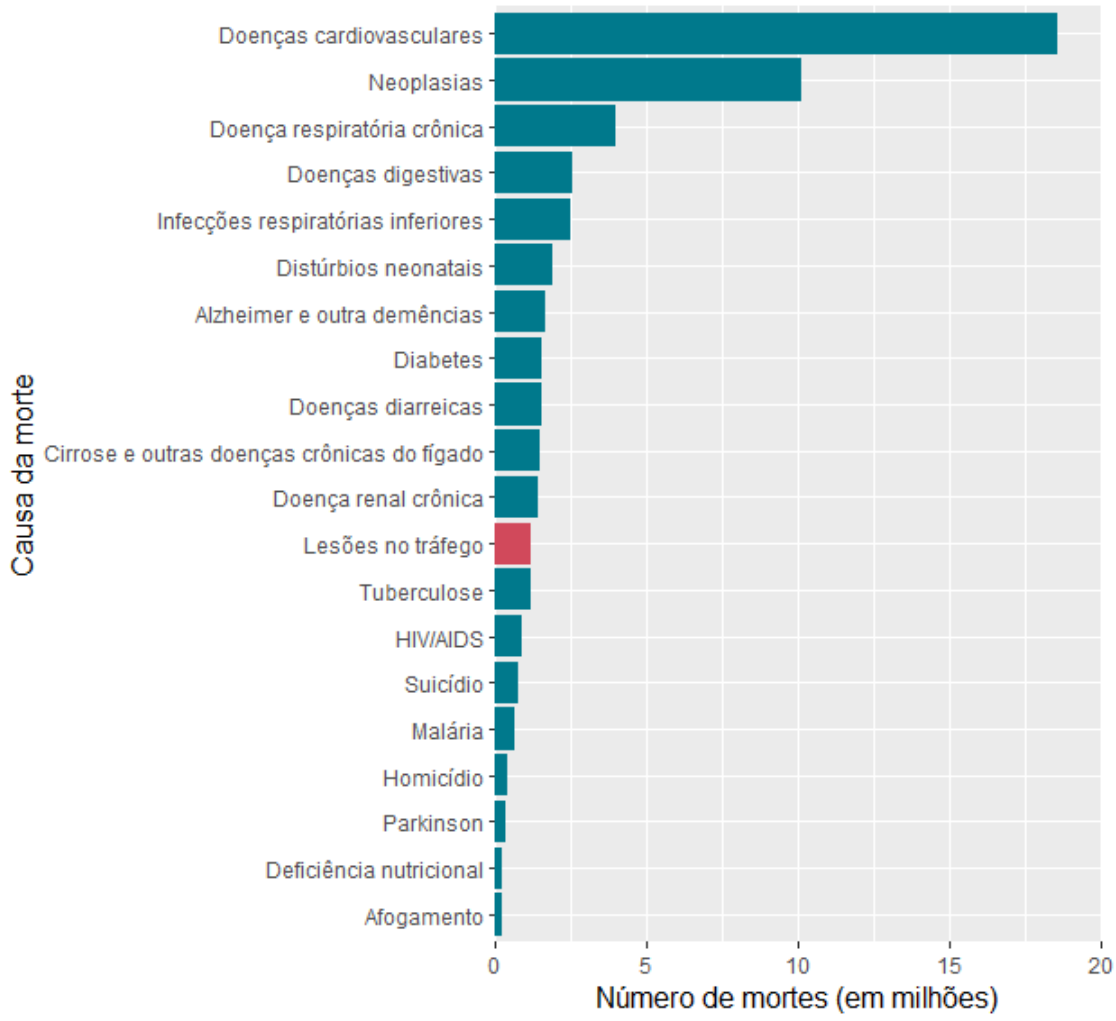

Figura 2.3: Vinte principais causas de morte - Mundo - 2019.
Fonte: Institute for Health Metrics and Evaluation (2019), com adaptações.

em 2019 com uma taxa ainda alta de aproximadamente 15 mortes por 100 mil habitantes, mas diferente do início da série, em 2010, quando eram observadas 21 mortes nas vias a cada 100 mil habitantes. Em detrimento a esses países, uma localidade exemplar nesse combate é a da Suécia, que conseguiu manter até 2018 uma taxa próxima a 3 mortes a cada 100 mil habitantes. Vale lembrar que a Suécia é a precursora do Visão Zero, iniciado em 1994 (Johansson, 2009).

Motivado pelos dados aqui apresentados, são buscadas técnicas para uma modelagem mais aprimorada de acidentes de trânsito fatais ou com vítimas, entendendo as variáveis explicati-

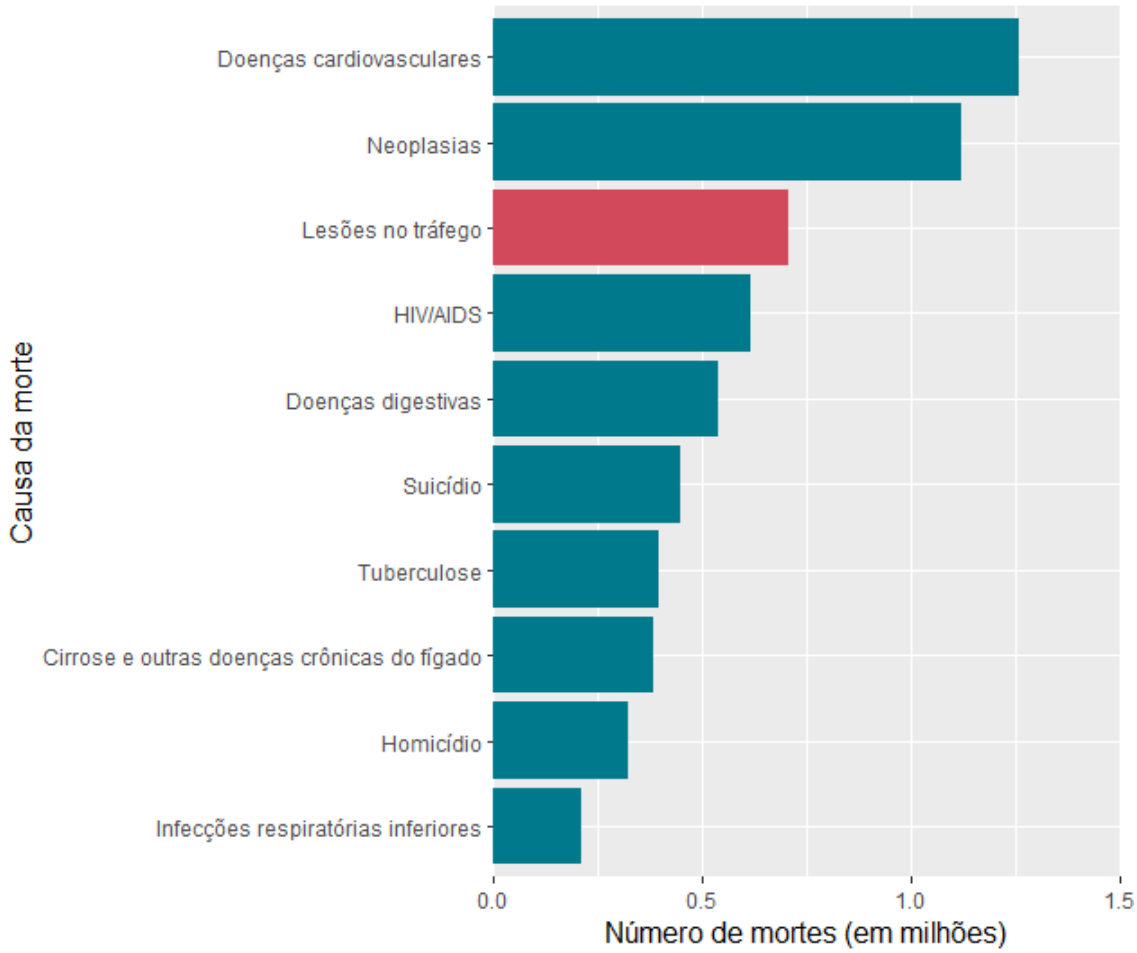

Figura 2.4: Dez principais causas de morte entre pessoas com 5 a 49 anos - Mundo - 2019.
Fonte: Institute for Health Metrics and Evaluation (2019), com adaptações.

vas que influenciam essas ocorrências, em busca de reduzir as consequências dos acidentes de trânsito.

Assim como nos modelos relativos à quantidade de acidentes de um modo geral, sem levar em consideração a presença de vítimas, os modelos que estudam a mortalidade nas vias contam com um arcabouço de informações previamente usadas, indicando variáveis conhecidamente importantes para esses problemas, como o excesso de velocidade, violação de regras de trânsito e falta do uso de cinto de segurança (Valent et al., 2002; Sivak et al., 2010; Siskind et al., 2011). Outros fatores como uma direção agressiva, sem carteira de motorista e com distrações

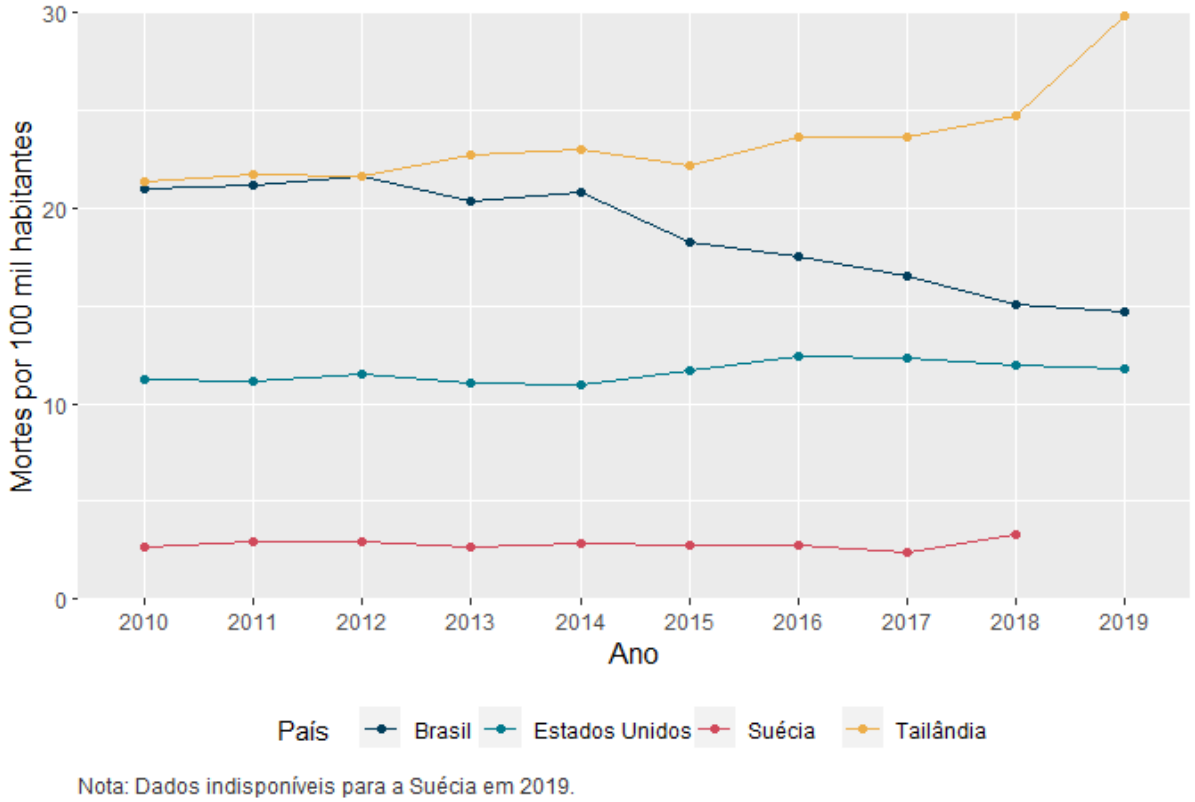

Figura 2.5: Mortalidade no trânsito - Brasil, Suécia, Tailândia e Estados Unidos - 2010 a 2019. Fonte: World Health Organization (2019), com adaptações.

durante a direção são vistas como importantes quando analisadas conjuntamente com a idade do motorista acidentado (Lambert-Bélanger et al., 2012; Hanna et al., 2012). Algumas outras características como o tipo de veículo (Fredette et al., 2008) e o horário da ocorrência (Arditi et al., 2007) também se mostram como fatores relevantes para o óbito nas vias, considerando estudos passados.

Para a modelagem, alguns estudos não se aprofundam na parte inferencial e destacam exploratoriamente a relação dos acidentes fatais com o local de ocorrência, como em Oris (2011); De Andrade et al. (2014) que utilizam estimativas de densidade *kernel*. Adentrando nos modelos estatísticos inferenciais, é frequente o uso de regressões logísticas, seja esta sem a utilização do fator espacial (Sivak et al., 2010; Siskind et al., 2011; Valent et al., 2002) ou com a localidade adicionada ao modelo, como apresentado em Hanna et al. (2012), por meio da regressão

logística condicional estratificada por localidade.

Alguns trabalhos utilizam como variável dependente a contagem de acidentes fatais como em Aguero-Valverde e Jovanis (2006), que considera para a modelagem a distribuição binomial negativa e uma abordagem Bayesiana, e em Rhee et al. (2016), com modelos espaciais econométricos.

Diante disso, algumas falhas metodológicas nos trabalhos supracitados são percebidas. Uma delas é desconsiderar a provável dependência espacial das ocorrências, quebrando um pressuposto fundamental que é a de independência das observações. Outra falha ocorre por utilizar a contagem de acidentes de trânsito com fatalidade ou com vítimas, uma vez que essa contagem é naturalmente influenciada pelo número de carros na localidade e não somente pela gravidade dos acidentes em determinado local, que é o fator que se busca entender aqui nesse estudo.

Por conta disso, a análise aqui construída busca uma melhor adaptação do dado para com sua distribuição real somando junto a isso a dependência espacial das ocorrências, sem desconsiderar os numerosos avanços, como a pré determinação de fatores fundamentais para a modelagem de acidentes com a presença de vítimas.

Capítulo 3

Regressão Beta Geograficamente Ponderada

3.1 Introdução

A regressão beta, desenvolvida por Ferrari e Cribari-Neto (2004), busca modelar dados advindos de taxas ou proporções, uma vez que a regressão linear tradicional não é mais adequada para esses problemas. Isso se deve a fatores como a restrição do suporte ao intervalo (0,1), a assimetria comum à esses dados e uma maior variação em torno do valor médio quando comparado a valores mais extremos da distribuição, caracterizando a heterocedasticidade (Dyke e Patterson, 1952).

A fim de incorporar uma abordagem espacial ao estudo de taxas e proporções, foi proposto por Da Silva e Lima (2017) a construção de um modelo de regressão para dados distribuídos no espaço e que, a princípio, indicam uma relação entre si, por conta da proximidade das ocorrências e, além disso, relações específicas a cada localidade de ocorrência. Para isso, deve ser considerado um modelo local, que leva em consideração tais fatores (Fotheringham et al., 2002).

Alguns ajustes à modelos de regressão com fins de examinar as relações locais são propos-

tos na literatura, como as técnicas de uso de funções spline (Wahba, 1990), regressão LOWESS (Cleveland, 1979) e regressão kernel (Cleveland e Devlin, 1988). Entretanto, estes modelos ainda desconsideram algumas variações espaciais importantes na relação entre a variável predita e suas preditoras (Fotheringham et al., 2002). Por conta disso, Brunsdon et al. (1996) desenvolveram o modelo de Regressão Geograficamente Ponderado (RGP), que visa capturar essa variação e, anos depois, como posto, Da Silva e Lima (2017) desenvolveram o modelo de Regressão Beta Geograficamente Ponderada (RBGP), que será aqui aplicado.

3.2 Caracterização da distribuição beta

Uma variável aleatória Y tem distribuição beta de parâmetros $\alpha > 0$ e $\beta > 0$ quando sua função densidade de probabilidade é dada por

$$f(y|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1},$$

com $0 < y < 1$ e $\Gamma(\cdot)$ sendo a função gama (Elderton, 1906). Os parâmetros α e β definem os diversos formatos que a distribuição beta pode ter, como pode ser visto na Figura 3.1. É possível obter uma distribuição em forma de J, de U, ou de J invertido (a), com diferentes simetrias (b) ou peso nas caudas da distribuição (c), ou até mesmo se ajustar a um comportamento linear (d). Se $\alpha = \beta$, tem-se uma distribuição simétrica em torno de $1/2$. Caso $\alpha < \beta$ a distribuição é assimétrica à direita e, se $\alpha > \beta$ tem-se a assimetria à esquerda. Ainda, quanto menor o valor dos parâmetros, maior será o peso das caudas da distribuição. No caso particular em que $\alpha = \beta = 1$, tem-se uma distribuição Uniforme no intervalo (0,1).

É conhecido que a média e a variância da distribuição beta são dadas, respectivamente, por

$$\mu = E(y) = \frac{\alpha}{\alpha + \beta},$$

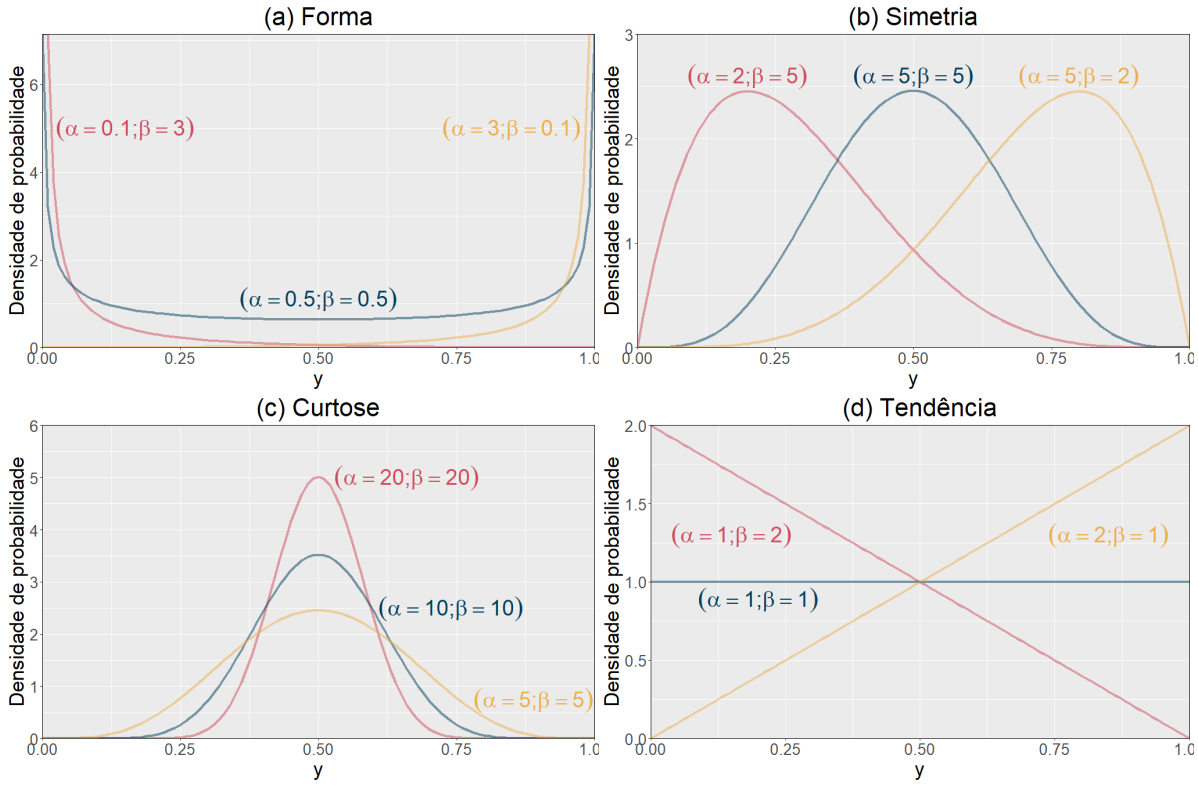

Figura 3.1: Distribuição Beta para diferentes valores dos parâmetros α e β.

e

$$\mathrm{Var}(y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Como a intenção é definir um modelo de regressão, torna-se mais interessante uma reparametrização da distribuição beta em função de sua média (μ) e que também considera um parâmetro para a precisão (ϕ) (Ferrari e Cribari-Neto, 2004).

3.3 Caracterização do modelo de regressão beta

Foi desenvolvido por Ferrari e Cribari-Neto (2004) um modelo adequado para situações em que o comportamento da variável resposta pode ser modelado como função de um conjunto de variáveis explicativas, como numa regressão tradicional porém, levando em consideração para a variável resposta a distribuição beta, citada acima, que restringe a análise ao intervalo contínuo

(0,1) e que possui uma grande flexibilidade para a modelagem.

Para isso, Ferrari e Cribari-Neto (2004) propuseram uma reparametrização considerando $\mu = \alpha/(\alpha + \beta)$ e $\phi = \alpha + \beta$, de modo que

$$\mathrm{E}(y) = \mu,$$

e

$$\mathrm{Var}(y) = \frac{V(\mu)}{1+\phi} = \frac{\mu(1-\mu)}{1+\phi}.$$

Assim, a distribuição beta reparametrizada fica na forma

$$f(y|\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \tag{3.1}$$

com $0 < \mu < 1$ e $\phi > 0$.

É natural que a flexibilidade apresentada na Figura 3.1 se mantenha porém, agora variando com relação aos parâmetros μ e ϕ, como na Figura 3.2.

Desta forma, se $\mu = 1/2$ tem-se a simetria encontrada quando anteriormente $\alpha = \beta$. Adicionalmente, nota-se que o parâmetro ϕ interfere apenas na precisão de tal forma que, para μ fixo, quanto maior ϕ, menor a variância da distribuição.

Considerando $y_1, y_2, \ldots, y_n$ variáveis aleatórias independentes que seguem uma distribuição beta reparametrizada (3.1) com média μ_t e precisão ϕ, para $t = 1, 2, \ldots, n$, o modelo pode ser obtido então assumindo que a média μ_t assume a estrutura de regressão

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = \eta_t, \tag{3.2}$$

ou seja, $\mu_t = g^{-1}(\eta_t)$, sendo:

- $x_{t1}, \ldots, x_{tk}$ os valores fixados para as k variáveis explicativas do modelo;

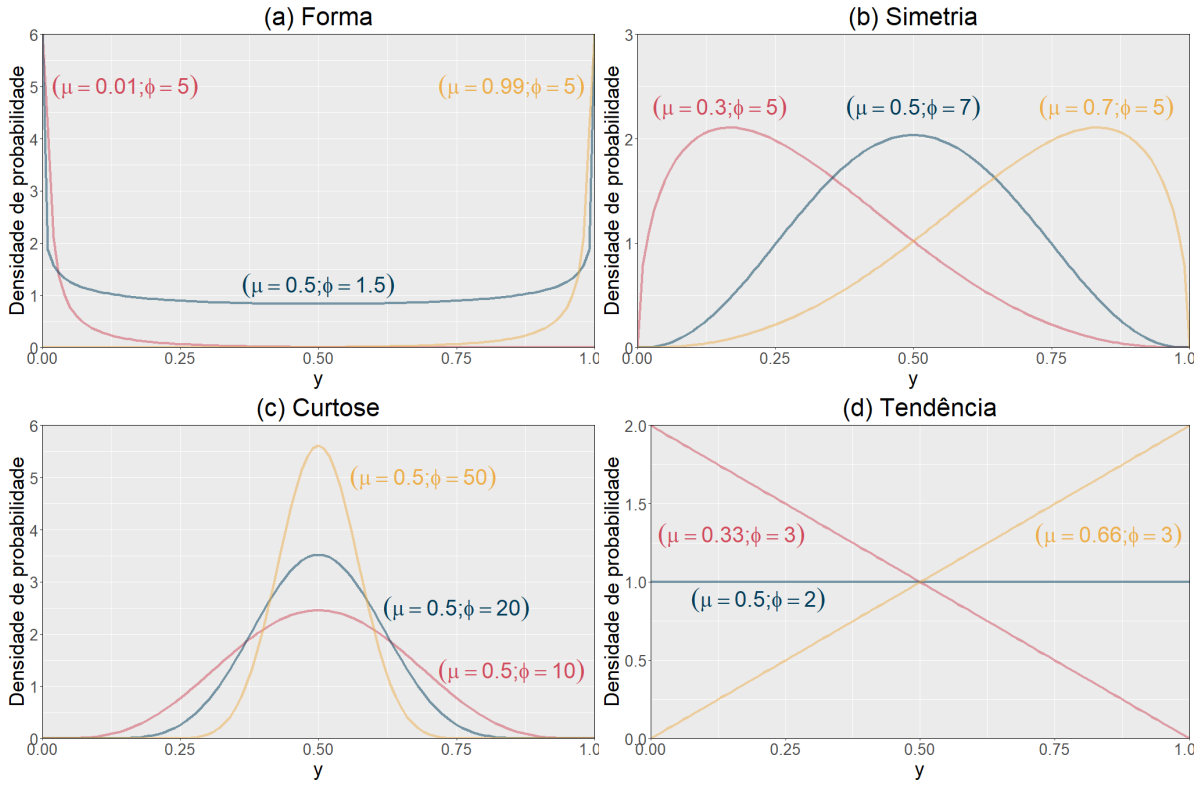

Figura 3.2: Distribuição Beta em função de μ e ϕ para diferentes valores de parâmetros.

- β_k o coeficiente da regressão para a variável x_k, $k < n$;
- $g(\cdot)$ uma função estritamente monótona e duplamente diferenciável, que associa o intervalo (0,1) à reta real da forma $\mu_t \in (0,1) \implies g(\mu_t) \in \mathbb{R}$, chamada função de ligação, sendo as mais comuns:
  - logito: $g(\mu_t) = \log\left(\frac{\mu_t}{1-\mu_t}\right)$;
  - probito: $g(\mu_t) = \Phi^{-1}(\mu_t)$, onde $\Phi(\cdot)$ é a função de distribuição acumulada da distribuição normal padrão;
  - log-log: $g(\mu_t) = -\log(-\log(\mu_t))$;
  - complemento log-log: $g(\mu_t) = \log(-\log(1-\mu_t))$.

A fim de estimar os parâmetros do modelo é utilizado o método de máxima verosimilhança,

e a partir da diferenciação do logaritmo da função de verossimilhança

$$\ell(\boldsymbol{\beta}, \phi) = \sum_{t=1}^{n} \ell_t(\mu_t, \phi)$$

com relação a $\beta_i, i = 1, \ldots, k$ e ϕ, é obtida a função escore

$$\begin{aligned} \boldsymbol{U}(\boldsymbol{\beta}, \phi) &= \begin{pmatrix} U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \phi) \\ U_{\phi}(\boldsymbol{\beta}, \phi) \end{pmatrix} = \begin{pmatrix} \frac{\partial \ell(\beta, \phi)}{\partial \boldsymbol{\beta}} \\ \frac{\partial \ell(\beta, \phi)}{\partial \phi} \end{pmatrix} = \\ &= \begin{pmatrix} \phi \boldsymbol{X}^\top \boldsymbol{T}(\boldsymbol{y^*} - \boldsymbol{\mu^*}) \\ \sum_{t=1}^{n} \{\mu_t(y_t^* - \mu_t^*) + log(1 - y_t) - \psi((1 - \mu_t)\phi) + \psi(\phi)\} \end{pmatrix} \end{aligned}$$

com $\mathbf{X}$ sendo a matriz $n \times k$ com os valores das covariáveis, $\mathbf{T} = \mathrm{diag}(1/g'(\mu_1), \ldots, 1/g'(\mu_n))$, e os vetores $\boldsymbol{y^*}$ e $\boldsymbol{\mu^*}$ sendo compostos, respectivamente, pelos elementos $y_t^* = \log(y_t/(1 - y_t))$ e $\mu_t^* = \psi(\mu_t\phi) - \psi((1 - \mu_t)\phi)$, com $\psi(\cdot)$ a função digama.

Entretanto, como os estimadores $\hat{\boldsymbol{\beta}}$ e $\hat{\phi}$ não possuem forma fechada é necessária a maximização numérica do logaritmo da função de verossimilhança a partir de algum algoritmo de otimização não-linear, como Newton (Simpson, 1740) ou Quasi-Newton(Fletcher e Powell, 1963), por exemplo.

Como valor inicial para esses algoritmos iterativos, Ferrari e Cribari-Neto (2004) sugerem os valores estimados por uma regressão linear clássica dos valores transformados ($\check{y} = g(y_t)$), ou seja, $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\check{\mathbf{y}}$ para $\boldsymbol{\beta}$ e para ϕ,

$$\frac{1}{n}\sum_{t=1}^{n} \frac{\check{\mu}_t(1 - \check{\mu}_t)}{\check{\sigma}_t^2} - 1$$

onde $\check{\mu}_t = g^{-1}(\mathbf{x}_t^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\check{\mathbf{y}})$ e $\check{\sigma}_t^2 = \frac{\check{\mathbf{e}}^\top\check{\mathbf{e}}}{(n-k)g'(\check{\mu}_t)^2}$, com $\boldsymbol{x}_j$ sendo a j-ésima linha da matriz $\boldsymbol{X}$ e $\check{\mathbf{e}}$ o vetor de resíduos ordinários da regressão linear clássica.

Com isso, sob as devidas condições de regularidade, para os estimadores de máxima veros-

similhança tem-se de forma assintótica

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\phi}} \end{pmatrix} \overset{a}{\sim} \mathcal{N}_{k+1}\left(\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\phi} \end{pmatrix}, \mathbf{K^{-1}} \right), \tag{3.3}$$

em que

$$\mathbf{K^{-1}} = \begin{pmatrix} K^{\beta\beta} & K^{\beta\phi} \\ K^{\phi\beta} & K^{\phi\phi} \end{pmatrix}, \tag{3.4}$$

com

$$K^{\beta\beta} = \frac{1}{\phi}\left(\mathbf{X}^\top\mathbf{Z}\mathbf{X}\right)^{-1}\left[\mathbf{I}_k + \frac{\mathbf{X}^\top\mathbf{T}\mathbf{c}\mathbf{c}^\top\mathbf{T}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{Z}\mathbf{X}\right)^{-1}}{\gamma\phi}\right],$$

$$K^{\beta\phi} = \left(K^{\phi\beta}\right)^\top = -\frac{1}{\gamma\phi}\left(\mathbf{X}^\top\mathbf{Z}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{T}\mathbf{c}$$

e

$$K^{\phi\phi} = \gamma^{-1}$$

em que

- $\gamma = tr(\mathbf{D}) - \phi^{-1}\mathbf{c}^\top\mathbf{T}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{Z}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{T}\mathbf{c}$;
- $\mathbf{Z}$ é uma matriz diagonal de tamanho n formada pelos elementos $z_t = \phi\{\psi'(\mu_t\phi)+\psi'((1-\mu_t)\phi)\}\frac{1}{g'(\mu_t)^2}$ em que $\psi'(\cdot)$ é a função trigama;
- $\mathbf{D}$ é uma uma matriz diagonal de tamanho n formada pelos elementos $d_t = \psi'(\mu_t\phi)\mu_t^2 + \psi'((1-\mu_t)\phi)(1-\mu_t)^2 - \psi'(\phi)$;
- $\mathbf{c}$ é um vetor de tamanho n formado pelos elementos $c_t = \phi\{\psi'(\mu_t\phi)\mu_t - \psi'((1-\mu_t)\phi)(1-\mu_t)\}$;
- $\mathbf{I}_k$ é a matriz identidade de ordem k.

Assumindo a análise de resíduos e a qualidade do ajuste como satisfatórias, é possível considerar a construção de intervalos de confiança para os parâmetros, conforme indicado por Ferrari

e Cribari-Neto (2004). Desta forma, para os coeficientes de regressão $\beta_j, j = 1, \cdots, k$, tem-se o intervalo de $(1-\alpha)100\%$ de confiança aproximado

$$\mathrm{IC}\left(\beta_j; (1-\alpha)\,100\%\right) = \hat{\beta}_j \pm \Phi^{-1}\left(1-\alpha/2\right) \times \mathrm{EP}(\hat{\beta}_j),$$

em que $EP(\hat{\beta}_j)$ é o erro padrão assintótico do estimador de máxima verossimilhança de $\hat{\beta}_j$ obtido a partir da inversa da matriz de Informação de Fisher avaliada na estimativa de máxima verossimilhança. Para o parâmetro de precisão ϕ, construímos o intervalo aproximado

$$\mathrm{IC}(\phi; (1-\alpha)\,100\%) = \hat{\phi} \pm \Phi^{-1}\left(1-\alpha/2\right) \times \mathrm{EP}(\hat{\phi})$$

em que $\mathrm{EP}(\hat{\phi}) = \hat{\gamma}^{-1/2}$.

Para verificar a qualidade do modelo ajustado, Ferrari e Cribari-Neto (2004) sugerem a utilização de uma medida denominada de pseudo R^2 (R^2_p), calculada com base no quadrado da correlação amostral entre $\hat{\eta}$ e $g(y)$, sendo $0 \leq R^2_p \leq 1$. Quanto maior o valor de R^2_p, maior a explicabilidade do modelo com relação à variação dos dados da resposta. Outras medidas estatísticas para a validação de modelos podem ser utilizadas, como o *deviance* (Spiegelhalter et al., 2002).

Para a análise de resíduos, Ferrari e Cribari-Neto (2004) recomendam o uso dos resíduos padronizados, da forma

$$r_t = \frac{y_t - \hat{\mu}_t}{\sqrt{\widehat{\mathrm{Var}}(y_t)}}, \tag{3.5}$$

com $\hat{\mu}_t = g^{-1}(\mathbf{x}_t^\top \hat{\boldsymbol{\beta}})$ e $\widehat{\mathrm{Var}}(y_t) = (\hat{\mu}_t(1-\hat{\mu}_t))/(1+\hat{\phi})$.

Algumas outras medidas podem ser construídas baseando-se na matriz de influência, ou matriz *hat*, que é obtida por

$$\mathbf{S} = \mathbf{Z}^{1/2}\mathbf{X}(\mathbf{X}^\top\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Z}^{1/2}. \tag{3.6}$$

Outro resíduo comumente utilizado na regressão beta, proposto por Espinheira et al. (2008), é o resíduo ponderado padronizado 2 (RPP2), estimado da forma

$$r_t^{ww} = \frac{y_t^* - \hat{\mu}_t^*}{\sqrt{\nu_t(1 - s_{tt})}}, \tag{3.7}$$

com $y_t^* = \log(y_t/(1-y_t))$, $\hat{\mu}_t^* = \psi(\hat{\mu}_t\hat{\phi}) - \psi((1-\hat{\mu}_t)\hat{\phi})$, $\nu_t = \psi'(\hat{\mu}_t\hat{\phi}) + \psi'((1-\hat{\mu}_t)\hat{\phi})$ e s_{tt} o t-ésimo elemento da diagonal principal da matriz $\mathbf{S}$ apresentada na Equação 3.6.

Um outro tipo de resíduo, proposto por Dunn e Smyth (1996) e aplicado à regressão beta por Pereira (2019) é o resíduo quantílico, definido por

$$r_t^q = \Phi^{-1}\left(F(y_t; \hat{\mu}_t, \hat{\phi})\right), \tag{3.8}$$

onde $\Phi(\cdot)$ e $F(\cdot)$ são as funções de distribuição acumulada da normal padrão e da distribuição beta, respectivamente.

Considerando o módulo desses resíduos é possível construir um gráfico de probabilidades meio-normal com o envelope simulado (Atkinson, 1981), que auxiliará na avaliação da qualidade do ajuste do modelo para com a distribuição proposta, no caso, a distribuição beta.

3.4 Caracterização da regressão beta geograficamente ponderada

Alguns trabalhos, com o intuito de adequar a modelagem local proposta por Fotheringham et al. (2002) à diferentes distribuições, foram desenvolvidos previamente, como os modelos logísticos (Albuquerque et al., 2017), Poisson (Nakaya et al., 2005), binomial negativo (Da Silva e Rodrigues, 2014) e beta (Da Silva e Lima, 2017). Esse último foi desenvolvido para tratar de taxas e proporções levando em conta a dependência espacial, considerando a média da variável resposta na localização j similar ao apresentado na Equação (3.2), exceto pela adição da

coordenada geográfica, como

$$g(\mu_j) = \eta_j = \sum_k \beta_k(u_j, v_j)x_{jk} = \sum_k \beta_{k(j)}x_{jk},$$

com $= 1, \ldots, n$, $g(\cdot)$ a função de ligação, $\beta_k(u_j, v_j) = \beta_{k(j)}$ o parâmetro para a k-ésima variável explicativa como função da j-ésima observação obtida no ponto de coordenadas (u_j, v_j), x_{jk} o valor observado para a k-ésima variável explicativa no ponto j e ε_j o erro para a j-ésima observação. Segundo Da Silva e Rodrigues (2014), é possível estimar os parâmetros de regressão para qualquer ponto i porém, as médias estimadas são calculadas apenas para os pontos j observados onde o valor x_{jk} é conhecido.

De acordo com Da Silva e Lima (2017), os parâmetros β_k podem ser estimados para o ponto i isto é, $\beta_i = \beta_i(u_i, v_i)$, usando o logaritmo da função de verossimilhança local

$$\ell(\boldsymbol{\beta_i}, \phi) = \sum_{j=1}^{n} \ell_j(\mu_j(\boldsymbol{\beta}(i)), \phi)w_{ij}, \tag{3.9}$$

com w_{ij} sendo os elementos de uma matriz espacial de pesos $\mathbf{W}$ que será descrita em breve.

Além disso, a função score do modelo para cada local i é representada por

$$\begin{aligned}\boldsymbol{U}(\boldsymbol{\beta_i}, \phi) &= \begin{pmatrix} U_{\boldsymbol{\beta_i}}(\boldsymbol{\beta_i}, \phi) \\ U_{\phi}(\boldsymbol{\beta_i}, \phi) \end{pmatrix} = \begin{pmatrix} \frac{\partial \ell(\boldsymbol{\beta_i}, \phi)}{\partial \boldsymbol{\beta_i}} \\ \frac{\partial \ell(\boldsymbol{\beta_i}, \phi)}{\partial \phi} \end{pmatrix} = \\ &= \begin{pmatrix} \sum_{j=1}^{n} \left\{ \phi(y_j^* - \mu_{j(i)}^*) \frac{1}{g'(\mu_{j(i)})} \boldsymbol{x}_j \right\} w_{ij} \\ \sum_{j=1}^{n} \left\{ \mu_{j(i)}(y_j^* - \mu_{j(i)}^*) + \log(1 - y_j) - \psi((1 - \mu_{j(i)})\phi) + \psi(\phi) \right\} w_{ij} \end{pmatrix},\end{aligned} \tag{3.10}$$

com $g'(\cdot)$ a derivada primeira da função de ligação, $y_j^* = \log(y_j/(1 - y_j))$, $\mu_{j(i)}^* = \psi(\mu_{j(i)}\phi) - \psi((1 - \mu_{j(i)})\phi)$.

Assim como no modelo de regressão beta, não existe uma forma fechada para a estimação

dos parâmetros $\boldsymbol{\beta}_k$ e ϕ, sendo necessário o uso de métodos de maximização numérica do logaritmo da função de verossimilhança local. Para essas otimizações Da Silva e Lima (2017) recomendam utilizar como ponto de partida para o algoritmo adaptações dos valores inicial da regressão beta, considerando agora uma matriz espacial de pesos $\boldsymbol{W}_i$, com base nas distâncias entre o local estimado e todos os pontos observados.

Os valores iniciais do vetor de parâmetros $\boldsymbol{\beta}_{0i}$ são estimados a partir da regressão geograficamente ponderada (Fotheringham et al., 2002) considerando $\check{y}_i = g(y_i)$, ou seja,

$$\boldsymbol{\beta}_{0i} = (\boldsymbol{X}^\top \boldsymbol{W_i} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W_i} \check{\boldsymbol{y}}, \tag{3.11}$$

com $\boldsymbol{W_i}$ uma matriz diagonal com os pesos w_{ij} podendo ser definidos, segundo Fotheringham et al. (2002) de acordo com o parâmetro de suavização ou, em inglês, "*bandwidth*", de vizinho mais próximo, usando o kernel adaptável biquadrático, da forma:

$$w_{ij} = \begin{cases} \left[1 - (d_{ij}/b)^2\right]^2, & \text{caso } j \text{ seja um dos } n\text{-ésimos vizinhos mais próximos de } i. \\ 0, & \text{caso contrário.} \end{cases}$$

ou de acordo com o parâmetro de suavização de distância, usando a função kernel Gaussiana (Fotheringham et al., 2002):

$$w_{ij} = \exp\left\{-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right\}.$$

Em ambos os casos, o valor ótimo para o parâmetro de suavização pode ser encontrado otimizando o AICc, como mostrado em Nakaya et al. (2005).

Para o valor inicial do parâmetro de precisão para o local (u_i, v_i), tem-se

$$\phi_{0i} = \frac{1}{n}\sum_{j=1}^{n} \frac{\check{\mu}_{0j}(1-\check{\mu}_{0j})}{\check{\sigma}^2_{0j}} - 1, \tag{3.12}$$

com $\check{\mu}_{ij} = g^{-1}(\boldsymbol{x}_j^\top(\boldsymbol{X}^\top \boldsymbol{W_i} \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W_i}\check{\boldsymbol{y}})$, sendo $\boldsymbol{x}_j$ a j-ésima linha da matriz $\boldsymbol{X}$ e $\check{\sigma}^2_{0j} = \frac{\check{\boldsymbol{e}}^\top \check{\boldsymbol{e}}}{(n-p_e)g'(\check{\mu}_{0j})^2}$, onde $\check{\boldsymbol{e}}$ é o vetor de resíduos ordinários da RGP considerando $\check{\boldsymbol{y}}$ e $p_e = 2\nu_1 - \nu_2$ o número efetivo de parâmetro do modelo RGP, com ν_1 o traço da matriz $\boldsymbol{S}$ e ν_2 o traço de $\boldsymbol{S}^\top \boldsymbol{S}$ (Fotheringham et al., 2002).

A matriz de informação de Fisher é obtida utilizando as derivadas parciais da função score, dada na Equação (3.10), ficando da forma

$$\boldsymbol{K} = \begin{pmatrix} K_{\beta\beta} & K_{\beta\phi} \\ K_{\phi\beta} & K_{\phi\phi} \end{pmatrix}, \tag{3.13}$$

com $K_{\beta\beta} = \boldsymbol{X}^\top \boldsymbol{\Phi} \boldsymbol{W} \boldsymbol{Z} \boldsymbol{X}$, $K_{\beta\phi} = K_{\phi\beta}^\top = \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{T} \boldsymbol{c}$ e $K_{\phi\phi} = \mathrm{tr}(\boldsymbol{W}\boldsymbol{D})$, sendo $\boldsymbol{\Phi}$, $\boldsymbol{Z}$, $\boldsymbol{T}$ e $\boldsymbol{D}$ matrizes diagonais com os respectivos elementos podendo ser calculados por:

- Para $\boldsymbol{\Phi}$: $\phi_i = \hat{\phi}_i$;
- Para $\boldsymbol{Z}$: $z_i = \hat{\phi}_i\{\psi'(\hat{\mu}_i\hat{\phi}_i) + \psi'((1-\hat{\mu}_i)\hat{\phi}_i)\}\frac{1}{g'(\hat{\mu}_i)^2}$;
- Para $\boldsymbol{T}$: $t_i = \frac{1}{g'(\hat{\mu}_i)}$; e
- Para $\boldsymbol{D}$: $d_i = \psi'(\hat{\mu}_i\hat{\phi}_i)\hat{\mu}_i^2 + \psi'((1-\hat{\mu}_i)\hat{\phi}_i)(1-\hat{\mu}_i)^2 - \psi'(\hat{\phi}_i)$.

e $\boldsymbol{c}$ o vetor de dimensão n com os elementos $c_i = \hat{\phi}_i\{\psi'(\hat{\mu}_i\hat{\phi}_i)\hat{\mu}_i - \psi'((1-\hat{\mu}_i)\hat{\phi}_i)(1-\hat{\mu}_i)\}$.

Vale ressaltar que, caso todos os pesos sejam iguais a 1, sendo então $\mathbf{W} = \mathbf{I}_k$, a matriz $\mathbf{K}$ retorna ao valores obtidos para a Equação (3.4), assim como os valores iniciais utilizados para a convergência das estimativas (Equações (3.11) e (3.12)).

Sob as devidas condições de regularidade, tem-se a distribuição assintótica dos parâmetros estimados $\hat{\boldsymbol{\beta}}$ e $\hat{\phi}$ como em (3.3), utilizando como variância os elementos que ocupam a diagonal principal da matriz de Informação de Fisher, dada na Equação (3.13).

Os procedimentos para inferência dos parâmetros e diagnóstico do modelo são análogos aos apresentados na Seção 3.3 porém, agora a matriz $\mathbf{S}$, dada anteriormente na Equação (3.6), é da

forma

$$\mathbf{S} = \mathbf{Z}^{1/2}\mathbf{X}(\mathbf{X}^\top\mathbf{W}\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{Z}^{1/2}. \tag{3.14}$$

Essa matriz é fundamental pois seu traço mensura o número real de parâmetros estimados pelo modelo, ou seja, $\nu_1 = \mathrm{tr}(\mathbf{S})$.

Na Regressão Geograficamente Ponderada, para cada estimativa de beta global, têm-se uma estimativa de betas locais para cada ponto da amostra, ou seja, n β_k estimados. A significância local para a estimativa do k-ésimo parâmetro no ponto i pode ser avaliada por meio do *pseudo* teste t, da forma:

$$t_k(u_i, v_i) = \frac{\hat{\beta}_{k(i)}}{EP\left[\hat{\beta}_{k(i)}\right]},$$

cuja distribuição é aproximadamente Normal padrão (Fotheringham et al., 2002).

Uma adaptação para esse teste é proposta por Da Silva e Fotheringham (2016), alterando o nível de significância α levando em consideração o número real de parâmetros:

$$\alpha = \frac{p}{p_e}\xi_m = \frac{\xi_m}{\frac{p_e}{p}}, \tag{3.15}$$

com $p_e = 2\mathrm{tr}(\mathbf{S}) - \mathrm{tr}(\mathbf{S}^\top\mathbf{S})$ o número efetivo de parâmetros independentes estimados na RGP e ξ_m representa o nível desejado de α desconsiderando a dependência espacial.

Capítulo 4

Materiais e Métodos

4.1 Introdução

O objetivo deste Capítulo é apresentar os materiais e métodos utilizados neste trabalho. Serão utilizados dados referentes a acidentes de trânsito ocorridos na cidade de Fortaleza-CE, entre os anos de 2009 e 2011. O objetivo do estudo aqui proposto é realizar a aplicação do modelo de Regressão Beta Geograficamente Ponderada (RBGP) à taxa de acidentes com vítimas. Além disso, será desenvolvido um pacote no *software* R a partir da macro SAS desenvolvida por Da Silva e Lima (2017), permitindo assim uma maior visibilidade à técnica, visto a ampla utilização do *software* R.

4.2 Materiais

O município de Fortaleza, no Ceará atualmente é dividido em 12 regionais, como apresentado na Figura 4.1, buscando a partir dessas divisões agregar bairros seguindo critérios como número de habitantes, afinidade socioeconômica e cultural entre esses bairros além da disponibilidade de equipamentos públicos (O Povo, 2021).

Porém, para o estudo aqui apresentado, foi utilizada uma base de dados georreferenciada dividida em 126 Zonas de Tráfego, com informações socioeconômicas e sobre o uso do solo

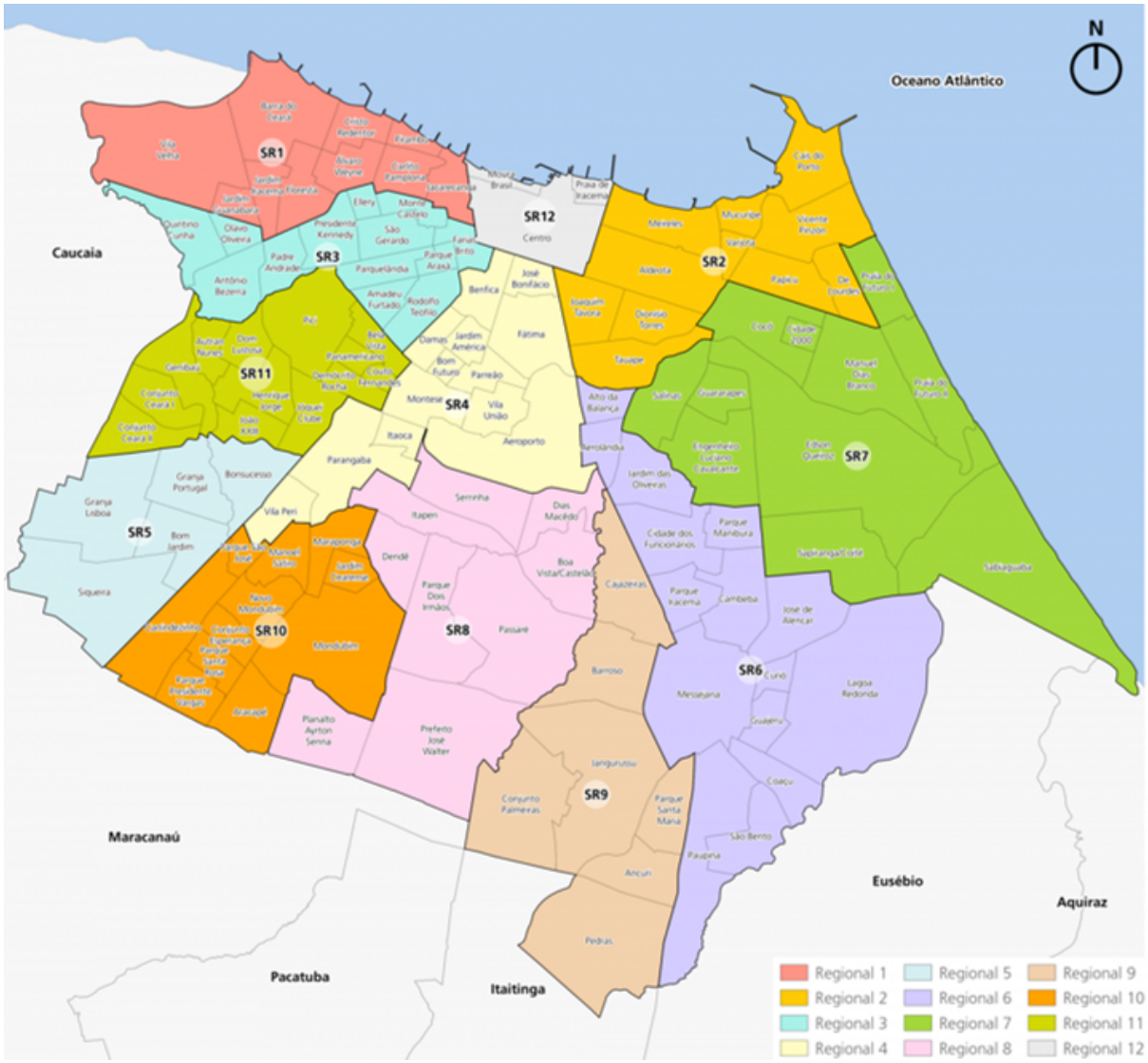

Figura 4.1: Divisões regionais do município de Fortaleza-CE.
Fonte: O Povo (2021), com adaptações.

advindas do Censo 2010 agregadas por essas ZT's. Junto a isso, o Sistema de Informações em Acidentes de Trânsito de Fortaleza (SIAT/FOR) disponibilizou variáveis da infraestrutura da rede, com a localização dos semáforos, os radares de velocidade e informações sobre acidentes ocorridos no período de 2009 a 2011, identificando a geolocalização desses acidentes, juntamente com a informação quanto a existência de vítimas na ocorrência.

As variáveis do conjunto de dados junto a algumas de suas medidas descritivas estão descri-

tas na Tabela 4.1. Houve uma separação dessas variáveis em 6 diferentes categorias, de acordo com sua característica. São elas:

- **Geral:** Informações de identificação da zona, sua localidade e seu tamanho.
- **Acidentes:** Contagem de acidentes totais e acidentes com vítimas.
- **Variável de exposição:** Tamanho das vias e população na ZT.
- **Características da rede:** Contagem de vias com e sem semáforo e equipamentos de fiscalização.
- **Características socioeconômicas:** Idade e renda familiar média das Zonas de Tráfego.
- **Uso do solo:** Área de uso do solo por comércio e por residências.

Tabela 4.1: Estatísticas descritivas das variáveis.

Categoria	Variável	Descrição	Média	Soma	Min.	Máx.	Desvio
Geral	ID	Código identificador da zona	-	-	-	-	-
	X	Coordenada Longitude em UTM (Zona 24S, Datum WGS 84)	-	-	-	-	-
	Y	Coordenada Latitude em UTM (Zona 24S, Datum WGS 84)	-	-	-	-	-
	N_ZONA	Número da Zona	-	-	-	-	-
	AREA_KM	Área da zona em km²	2,41	303,14	0,14	13,48	2,58
Acidentes	ACT	Nº de acidentes totais no período de 2009 a 2011	431,66	54389	6	2981	450,92
	ACV	Nº de acidentes com vítimas no período de 2009 a 2011	153,4	19328	4	829	119,94
Variável de exposição	EXT_TOT	Extensão total de vias na ZT (km)	33,26	4190,18	2,63	192,07	26,89
	POP_TOT	População total na ZT	19213,86	2420946	1183	115279	16846,7
Características da rede	DEN_I_SEM	N° de interseções semaforizadas por km de via	0,3	37,6	0	1,94	0,35
	DEN_I_NSEM	N° de interseções não semaforizadas por km de via	6,11	769,4	3,58	9,71	1,18
	D_EQUI_FE	N° de equipamentos de fiscalização eletrônica por km de via	0,07	8,41	0	0,39	0,07
Características socioeconômicas	P_0_17	Proporção de residentes entre 0 e 17 anos de idade	0,26	-	0,16	0,37	0,05
	P_18_64	Proporção de residentes entre 18 e 64 anos de idade	0,66	-	0,59	0,72	0,03
	P_M64	Proporção de residentes com 65 anos ou mais de idade	0,07	-	0,03	0,14	0,03
	P_D_A3SM	Proporção de domicílios com renda familiar até 3 salários mínimos	0,58	-	0,08	0,92	0,22
	P_D_M3SM	Proporção de domicílios com renda familiar acima de 3 salários mínimos	0,42	-	0,08	0,92	0,22
Uso do solo	URES_A	Uso do solo do tipo residencial (m²) por m² da ZT	0,24	30,18	0,00	1,55	0,24
	UCOPS_A	Uso do solo do tipo comercial e prestação de serviço (m²) por m² da ZT	0,09	11,68	0,00	0,74	0,1

Dentre as variáveis apresentadas, serão consideradas as que mais tem relação com a taxa de acidentes com vítimas, a ser calculada, e estas serão utilizadas no modelo a ser desenvolvido.

4.3 Métodos

O trabalho será dividido em duas etapas. A modelagem utilizando a regressão beta geograficamente ponderada e o desenvolvimento do pacote no *software* R.

4.3.1 Modelagem RBGP

Em busca de identificar fatores relacionados a acidentes de trânsito com vítimas juntamente com o vínculo espacial dessas ocorrências, será utilizado o modelo de regressão beta geograficamente ponderado.

Como visto no Capítulo 3, esse modelo utiliza como variável resposta uma taxa ou proporção. Essa taxa não é apresentada, necessitando assim a sua construção, dividindo a contagem de acidentes com vítimas no período estudado pelo número de acidentes totais (com vítimas ou somente danos materiais), ou seja,

$$\text{AC\_TX} = \frac{\text{ACV}}{\text{ACT}}. \tag{4.1}$$

Inicialmente será realizada uma análise exploratória dos dados, buscando os principais fatores para o modelo. Para evidenciar a suspeita de dependência espacial será utilizado a princípio o diagrama de Moran (Anselin, 1996), apresentado na forma de um mapa, que identifica observações com alto valor na variável estudada e estão cercados por outras observações que também possuem valor alto ou baixos valores cercados por outros valores baixos, ou ainda a coocorrência de valores altos e baixos num local. Também será utilizado o LISA (Anselin, 1995) que indica apenas os locais significativos do Mapa de Moran. Esses indicadores auxiliam na visualização da dependência espacial.

A fim de obter um valor numérico para essa possível dependência será utilizado o Índice de Moran, apresentado em Moran (1950) da forma $I = (\boldsymbol{z}^\top \boldsymbol{z})^{-1} \boldsymbol{z}^\top \boldsymbol{W} \boldsymbol{z}$, com $\boldsymbol{z} = (\boldsymbol{y} - \overline{\boldsymbol{y}})$ e $\boldsymbol{W}$ sendo uma matriz de dependência espacial $n \times n$, podendo este índice variar entre -1 e 1, onde -1 representa autocorrelação espacial negativa perfeita, 0 representa a ausência de correlação espacial e 1 representa a autocorrelação positiva perfeita. Assintoticamente, o I de Moran possui distribuição gaussiana, o que permite o desenvolvimento de testes para verificar a significância da autocorrelação espacial obtida.

Quanto aos fatores não relacionados ao espaço, uma matriz de correlação será desenvolvida

a fim de ter indícios sobre a relação da variável resposta com as possível variáveis explicativas para o modelo.

Assim, será realizada uma modelagem não-espacial, e após isso será adicionado o fator geográfico ao modelo. A ideia consiste em procurar fatores que podem não influenciar globalmente o modelo porém, quando analisada a característica em determinada área, existe a relação com a taxa de acidentes com vítimas.

O algoritmo *Golden Section Search* (GSS) (Luenberger e Ye, 1984) será utilizado a fim de identificar o parâmetro de suavização ótimo para a aplicação do modelo geograficamente ponderado. Esse valor define os vizinhos utilizados para o modelo, seja a partir do raio em quilômetros partindo do centroide do local de análise (fixo), ou pela contagem de vizinhos mais próximos de cada localização, podendo variar a distância para cada modelo (adaptável). Além disso, pode-se buscar a minimização da função de validação cruzada (CV) ou do AIC. Portanto, o modelo local poderá ser construído, identificando as variáveis que mais explicam a ocorrência de acidentes com vítimas, e quais destas também são influenciadas geograficamente.

Assim sendo, buscar-se-á comparar seis modelos, sendo três deles globais e três locais. Primeiramente, um modelo global considerando a distribuição normal para a variável resposta. O segundo modelo também utilizará o modelo de regressão normal global, mas para a variável resposta ajustada ao suporte nos números reais, $y_t^* = \log(y_t/(1-y_t)), t = 1, \ldots, n$, transformação que contornaria o problema dos limites definidos para proporções, o intervalo unitário. Por fim, o terceiro modelo global, considerando a distribuição beta.

Por último, serão desenvolvidos três modelos locais, sendo o primeiro considerando a distribuição normal (RGP) sem nenhuma transformação na variável resposta, um modelo ainda com distribuição normal porém agora transformado pela função logito (RGP-logito), e, o principal destes, tendo em vista a distribuição beta (RBGP).

Nesses modelos, para identificar a importância das variáveis em cada localidade, será considerada uma significância geral de 10%, que a depender do número de parâmetros estimados decai, conforme visto na Equação 3.15.

Tais modelos serão aplicados aos dados e sua qualidade será medida de acordo com as tradicionais métricas de qualidade de ajuste de modelos, além de verificar os valores previstos fora do intervalo $(0, 1)$, o que seria um erro grosseiro para o nosso problema.

4.3.2 Pacote R

Para o desenvolvimento do pacote no *software* R será considerado o algoritmo apresentado em Da Silva e Lima (2017) na linguagem SAS. A tradução do código ocorrerá utilizando simultaneamente os *softwares* R Studio, IDE para a linguagem R e o SAS 9.4 M7, a fim de comparar os resultados obtidos com ambos os *softwares*.

O pacote desenvolvido conta com a presença de uma função para identificar o parâmetro de suavização ótimo, via *Golden Section Search* (GSS) além da função para a realização do modelo RBGP.

No *software* R, os pacotes *"devtools"* e *"roxygen2"* foram utilizados, o primeiro com uma série de ferramentas para desenvolvimento de pacotes e o seguinte que auxilia na criação da documentação do pacote, onde se descreve cada argumento da função, juntamente com a finalidade de cada função.

Após o desenvolvimento do pacote, o mesmo será disponibilizado de forma pública na plataforma *GitHub*, tradicional plataforma de hospedagem e compartilhamento de códigos-fonte e após isso, o pacote será encaminhado para o *"The Comprehensive R Archive Network"* (CRAN), repositório de pacotes que são carregados de forma nativa na plataforma, possibilitando assim uma maior difusão da técnica aqui aplicada.

Capítulo 5

Resultados

5.1 Introdução

Esse Capítulo apresenta os resultados do método apresentado no Capítulo anterior, ou seja, ajustando um modelo de regressão beta geograficamente ponderado aos dados e comparando o seu desempenho com uma regressão geograficamente ponderada, e com uma regressão geograficamente ponderada, mas com a variável dependente modificada.

A próxima seção apresenta uma breve descrição da variável resposta estudada, a seção 5.3 estuda a correlação desta com as variáveis explicativas disponíveis, buscando identificar os principais fatores correlatos com os acidentes com vítimas e assim, ter informações iniciais para a definição do modelo. Após isso, os modelos propostos serão ajustados aos dados e comparados entre si.

5.2 Análise inicial da taxa de acidentes com vítimas

A variável resposta do estudo considerou o número de acidentes com vítimas em relação à todos os acidentes ocorridos entre 2009 e 2011 na cidade de Fortaleza-CE, como visto na Equação 4.1.

Como apresentado na Tabela 5.1, em média, entre as zonas de tráfego definidas, a taxa de

acidentes com vítimas é de 42,46%, com um desvio padrão de 13,62%. As ZT's com menor e maior taxa de acidentes com vítimas possuem, respectivamente, 17,08% e 71,43% das ocorrências. Além disso, 25% das zonas de tráfego de Fortaleza possuem uma taxa de acidentes com vítimas menor que 31,75%, 50% das zonas têm a taxa abaixo de 41,47% e 75% delas possuem uma taxa menor que 53,00%.

Tabela 5.1: Medidas descritivas da taxa de acidentes com vítimas.

Taxa de acidentes com vítimas			
Mínimo	$0,1708$	**N**	126
1º quartil	$0,3175$	**Média**	$0,4246$
Mediana	$0,4147$	**Desvio padrão**	$0,1362$
3º quartil	$0,5300$	**Curtose**	$-0,8720$
Máximo	$0,7143$	**Assimetria**	$0,2703$

A distribuição da variável resposta possui uma curtose de -0,87, indicando um peso irrisório nas caudas da distribuição. Além disso, foi encontrado um valor de 0,27 para o coeficiente de assimetria, indicando uma simetria considerável à distribuição, fazendo com que esta se aproxime bastante da distribuição Gaussiana, como pode ser visto na Figura 5.1. Ainda, analisando a adequabilidade das distribuições normal e beta aos dados, tem-se pelo teste de Kolmogorov-Smirnov (Conover, 1980) que as duas distribuições propostas se encaixam nos dados analisados, com uma maior evidência de ajuste da distribuição beta.

A fim de entender o comportamento da variável resposta no espaço estudado tem-se o mapa apresentado na Figura 5.2, onde as cores mais escuras indicam uma maior taxa de vitimas nos acidentes. Com essa análise, é visualmente perceptível a importância da adição do fator "espaço" ao modelo uma vez que as ZT's centrais se mostram com uma menor taxa de acidentes fatais quando comparado à zonas mais periféricas.

Vale ainda observar que algumas zonas aparecem em branco na Figura 5.2. Esses locais indicam áreas onde não existe tráfego. Por exemplo, de acordo com o mapa apresentado na Figura 4.1, o espaço em branco no centro do mapa é exatamente a área onde está localizado o aeroporto da cidade de Fortaleza.

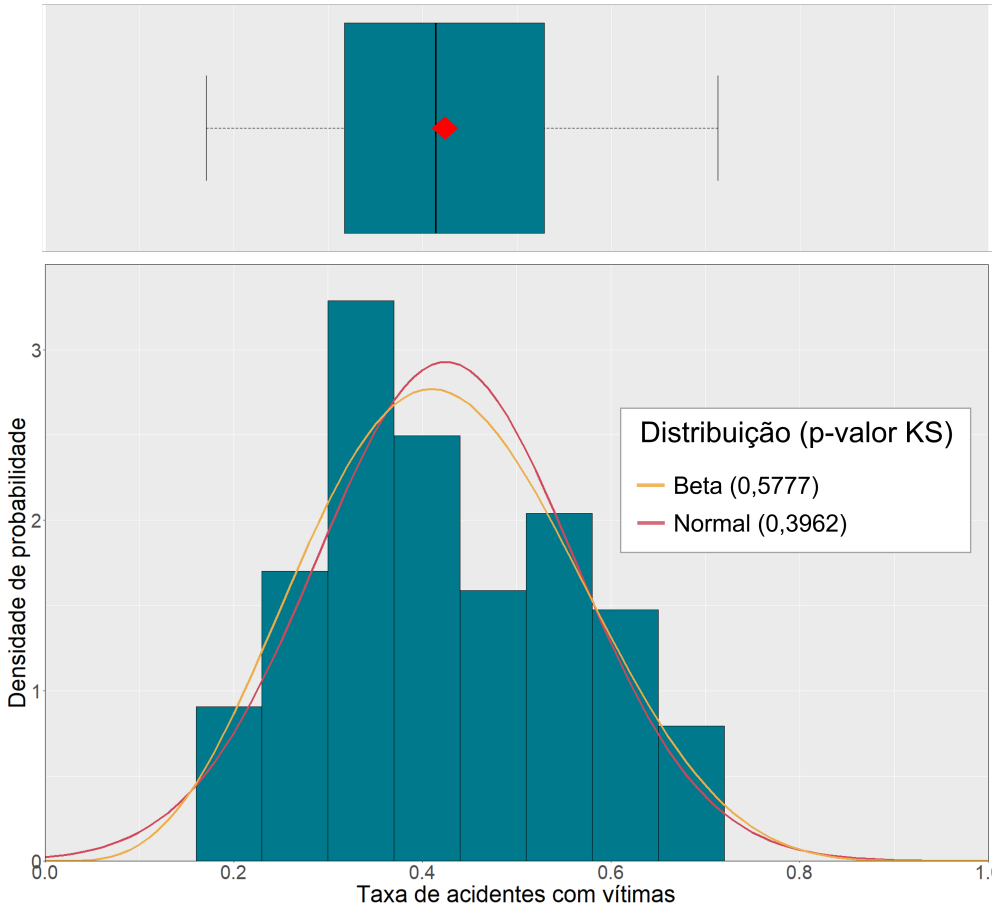

Figura 5.1: Distribuição da taxa de acidentes com vítimas.

Para ajudar a confirmação da análise espacial inicial pode ser construído um mapa a partir do diagrama de Moran (Figura 5.3), indicando a influência ou não das ZTs vizinhas em valores mais altos ou mais baixos para a taxa de acidentes com vítimas. Utilizando uma matriz de contiguidades $\boldsymbol{W}$, foi obtido $I = 0,7$, indicando uma correlação espacial consideravelmente forte para o evento estudado. Nota-se que, no geral, 82,5% das localidades com valor alto ou baixo possuem vizinhos com a mesma escala de valor, ou seja, considerando um valor alto tem-se em 38,1% dos casos valores altos como vizinhos e, para valores baixos, tem-se 44,4% de vizinhos com valores também baixos.

Analisando ainda a Figura 5.3, o Mapa LISA busca identificar os locais onde a influência dos vizinhos na ocorrência é significativa, a um nível de 95% de confiança. Considerando esse nível

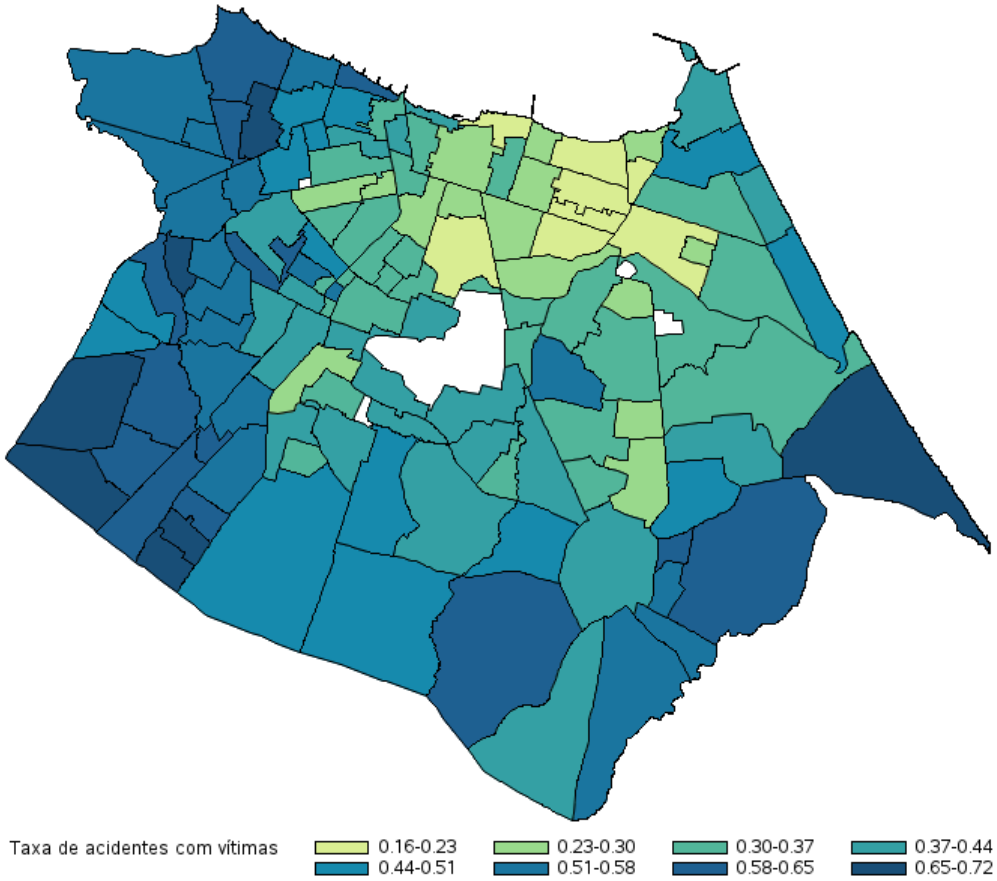

Figura 5.2: Distribuição espacial da taxa de acidentes com vítimas.

de significância, nota-se a influência de alguns locais mais periféricos da cidade para a elevação da taxa de acidentes fatais enquanto que, o centro da cidade indica a influência significativa de taxas menores de ocorrência.

5.3 Análise de correlação

Além da localização, outras variáveis devem ser estudadas para entender as ocorrências. Desta forma, são analisadas as correlações da taxa de acidentes com vítimas com as demais variáveis apresentadas, buscando conhecer quais são as variáveis mais correlatas ao evento e então construir os modelos propostos no Capítulo anterior.

Inicialmente, foram estudadas as variáveis utilizadas nos modelos por Gomes et al. (2017),

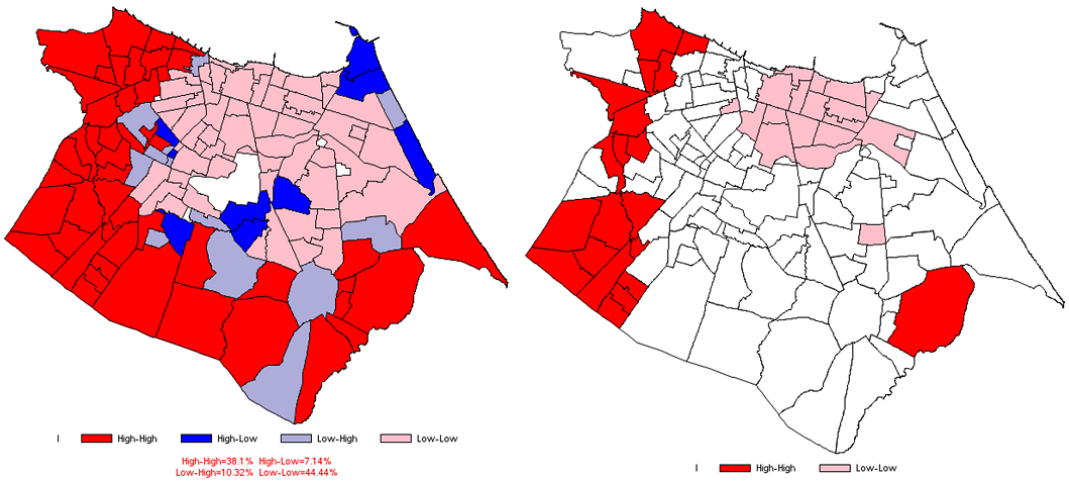

Figura 5.3: Mapa de Moran e Mapa LISA para a taxa de acidentes com vítimas.

que utilizou a mesma base de dados deste estudo porém, usando a contagem de acidentes com vítimas como variável resposta. Por conta dessa alteração, percebe-se uma diferença considerável nas medidas de correlação quando se usa a taxa ou a contagem de acidentes com vítimas (Tabela 5.2).

Tabela 5.2: Matriz de correlação com as variáveis utilizadas nos modelos apresentados em Gomes et al. (2017).

	AC_TX	AC_V09_11	LN(EXT_TOT)	P_M64	D_EQUI_FE	DEN_I_SEM
AC_TX	1	$-0,2091$	$0,2182$	$-0,6388$	$-0,3185$	$-0,5806$
AC_V09_11	$-0,2091$	1	$0,4673$	$0,3077$	$0,3623$	$0,4362$
LN(EXT_TOT)	$0,2182$	$0,4673$	1	$-0,2947$	$-0,0148$	$-0,3306$
P_M64	$-0,6388$	$0,3077$	$-0,2947$	1	$0,2856$	$0,6790$
D_EQUI_FE	$-0,3185$	$0,3623$	$-0,0148$	$0,2856$	1	$0,4028$
DEN_I_SEM	$-0,5806$	$0,4362$	$-0,3306$	$0,6790$	$0,4028$	1

Em vermelho, correlações não significativas considerando 95% de confiança.

Desta forma, a matriz de correlação com todas as variáveis disponibilizadas foi estudada a fim de selecionar as variáveis mais correlatas com a taxa de acidentes com vitimas. Além disso, tomou-se um cuidado especial em não selecionar para o modelo variáveis explicativas muito correlacionadas entre si, evitando o problema de multicolinearidade. A matriz completa se encontra no Anexo A e, a matriz de correlação com as variáveis selecionadas é apresentada na Tabela 5.3, juntamente com uma breve descrição das variáveis selecionadas (Tabela 5.4),

incluindo a visualização de como estas estão distribuídas e as relações entre elas (Figura 5.4).

Tabela 5.3: Matriz de correlação com as variáveis a serem utilizadas nos modelos.

	AC_TX	P_D_A3SM	DEN_I_SEM	D_EQUI_FE
AC_TX	1	$0,7956$	$-0,5806$	$-0,3185$
P_D_A3SM	$0,7956$	1	$-0,6307$	$-0,3297$
DEN_I_SEM	$-0,5806$	$-0,6307$	1	$0,4028$
D_EQUI_FE	$-0,3185$	$-0,3297$	$0,4028$	1

Tabela 5.4: Descrição das covariáveis selecionadas para o modelo.

Variável	**Descrição**	**Média**	**Soma**	**Min.**	**Máx.**	**Desvio**
P_D_A3SM	Proporção de domicílios com renda familiar até 3 salários mínimo	$0,58$	-	$0,08$	$0,92$	$0,22$
DEN_I_SEM	N° de interseções semaforizadas por km de via	$0,3$	$37,6$	0	$1,94$	$0,35$
D_EQUI_FE	N° de equipamentos de fiscalização eletrônica por km de via	$0,07$	$8,41$	0	$0,39$	$0,07$

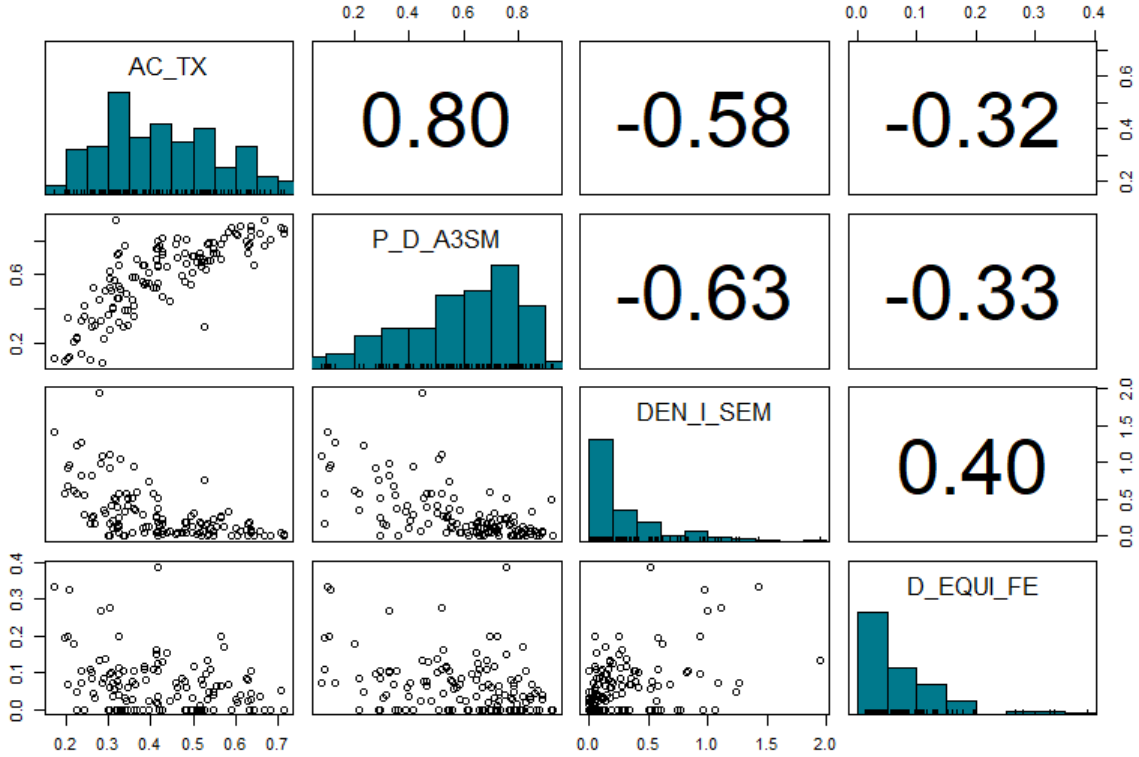

Figura 5.4: Distribuição e correlação das variáveis selecionadas para a modelagem.

Pode ser visto então que o fator mais associado à taxa de acidentes com vítimas é a proporção de domicílios com renda familiar até 3 salários mínimos, dando um indício de que em locais com menor concentração de renda ocorrem mais acidentes.

A maior correlação negativa em relação à variável resposta ocorre com o fator "número de interseções semaforizadas por quilômetro de via" ($\rho = -0.58$). Analisando exploratoriamente, quanto maior o número de cruzamentos com semáforos, menor a taxa de acidentes com vítimas.

A outra variável escolhida foi o número de equipamentos de fiscalização eletrônica por quilômetro de via, com correlação de $\rho = -0.32$ com a variável resposta.

Com as variáveis indicadas, agora é possível ajustar os modelos propostos anteriormente.

5.4 Ajuste de modelos globais

O primeiro modelo ajustado foi o modelo de regressão normal, sem qualquer transformação na variável resposta. Após isso, foram ajustados quatro modelos ainda utilizando a regressão normal porém com transformações segundo as funções de ligação logito, probito, log-log e complemento log-log na variável resposta. Comparando a explicabilidade dos modelos utilizando cada uma dessas transformações de acordo com o R^2 ajustado, os melhores resultados foram complemento log-log ($R^2_a = 0,6758$), logito ($R^2_a = 0,6563$), log-log ($R^2_a = 0,6249$) e, por último, probito ($R^2_a = 0,5752$).

Por conta de uma melhor interpretação dos parâmetros, foi selecionado para fins de comparação o modelo logito, que mesmo não sendo o modelo com melhor ajuste, mantém resultados satisfatórios.

Por fim, foram ajustados modelos de regressão beta global, utilizando o mesmo critério para a seleção do modelo porém, agora com a estatística Pseudo R^2 ajustado para identificar a função de ligação que atinge a melhor performance. Novamente, o melhor modelo foi ajustado utilizando a transformação complementar log-log porém, utilizando o mesmo argumento do modelo linear, será utilizado o modelo com a função logito, que neste caso também foi a que

produziu o segundo melhor ajuste aos dados dentre as funções testadas. Os três modelos globais foram ajustados e seus resultados são apresentados na Tabela 5.5.

Tabela 5.5: Resultados dos modelos globais.

Variável	**Regressão linear clássica**			**Regressão linear (logito)**			**Regressão beta (logito)**		
	Estimativa	t	p-valor	Estimativa	t	p-valor	Estimativa	t	p-valor
Intercepto	$0,1809$	$5,45$	$< 0,0001$	$-1,3721$	$-9,88$	$< 0,0001$	$-1,3532$	$-9,69$	$< 0,0001$
P_D_A3SM	$0,4487$	$10,13$	$< 0,0001$	$1,9412$	$10,46$	$< 0,0001$	$1,9165$	$10,29$	$< 0,0001$
DEN_I_SEM	$-0,0457$	$-1,65$	$0,1022$	$-0,2104$	$-1,81$	$0,0725$	$-0,2184$	$-1,85$	$0,0662$
D_EQUI_FE	$-0,0682$	$-0,62$	$0,5334$	$-0,3725$	$-0,81$	$0,4167$	$-0,3224$	$-0,71$	$0,4788$
Phi	-	-	-	-	-	-	$37,4164$	$8,04$	$< 0,0001$
R^2 ajustado*	$0,6357$			$0,6563$			$0,6535$		
AICc	$-267,7472$			$93,0940$			$-276,5724$		
Log-verossimilhança	$138,0389$			$-42,3817$			$143,5261$		

*Pseudo R^2 ajustado para a regressão beta.

As estimativas obtidas para o modelo de regressão linear clássica são bem distintas das obtidas pelos modelos com o uso da função logito porém, a interpretação desses parâmetros também é realizada de forma distinta.

Para a regressão linear clássica o aumento de 10 pontos percentuais na proporção de moradores com renda abaixo de 3 salários mínimos, mantendo as demais variáveis explicativas constantes, aumenta em média $4,5$ pontos percentuais na taxa de acidentes com vítimas. Se considerado o aumento de uma unidade no número de interseções semaforizadas por km de via, mantendo as mesmas condições, é esperada um decréscimo médio de $4,57$ pontos percentuais na variável resposta. Por fim, tem-se que o número de equipamentos de fiscalização eletrônica por km de via não é significativa para o modelo.

Por outro lado, para o modelo de regressão clássica com transformação logito, não é possível realizar uma interpretação direta dos parâmetros, uma vez que o valor encontrado indicaria quanto o valor de $\log(y_t/(1-y_t))$ aumentaria (ou diminuiria, em casos de coeficientes negativos).

Mesmo que estes tenham valores próximos aos obtidos com a regressão beta, como visto na Tabela 5.5, a interpretação dessas estimativas ocorre de maneira bem distinta.

Com isso, no modelo de regressão beta, onde a interpretação é realizada por meio da razão de chances, aumentando 10 pontos percentuais na proporção de domicílios com baixa renda

naquela ZT tem-se um aumento de $21,1\%$ na chance de ocorrência de acidentes com vítimas. A cada adição de semáforo em uma interseção por km de via tem-se uma queda de $19,6\%$ na chance de ocorrência e espera-se que essa chance caia em $27,6\%$ a cada adição de equipamento de fiscalização por km de via. Para todos esses casos considera-se que as outras variáveis explicativas não analisadas no momento se mantêm constantes.

Quanto à qualidade dos modelos apresentados, tem-se melhores métricas no modelo de regressão beta, considerando a análise do AICc e da função de log-verossimilhança. Contudo, essas métricas não são comparáveis, uma vez que tratam de modelos com diferentes técnicas, também chamados de "modelos não encaixados".

Se analisado o R^2 ajustado, medida que torna possível a comparação entre os modelos, tem-se uma melhor explicabilidade quando utilizado o modelo de regressão linear com transformação pela função logito. Entretanto, a diferença entre esse valor (65,6%) e o obtido com a regressão beta (65,4%) não parece justificar o uso do modelo transformado, uma vez que quando este é utilizado, perde-se o fator de interpretação dos parâmetros estimados, não sendo mais possível o uso da razão de chances.

Analisando a distribuição dos resíduos studentizados gerados pelos modelos com distribuição normal (Figura 5.5), notam-se fortes indícios de heterocedasticidade não capturada pelos modelos, uma vez que a variância dos resíduos parece aumentar com o aumento dos valores ajustados. O mesmo ocorre nos resíduos ponderados padronizados 2 (RPP2) do modelo de regressão beta.

Quando observados os gráficos de probabilidade meio-normal, observa-se um bom ajuste dos resíduos dentro dos envelopes em todos os modelos. Vale ressaltar ainda que para a regressão beta a adequação dos resíduos à distribuição gaussiana não é um pressuposto necessário.

Os valores preditos para o modelo de regressão linear utilizando a função logito aparentam estar fora do escopo $(0,1)$ porém, os valores obtidos ainda precisariam ser retransformados, considerando o inverso da função de ligação utilizada. Em nenhum dos três modelos foram obtidos valores ajustados para a resposta fora do intervalo $(0,1)$.

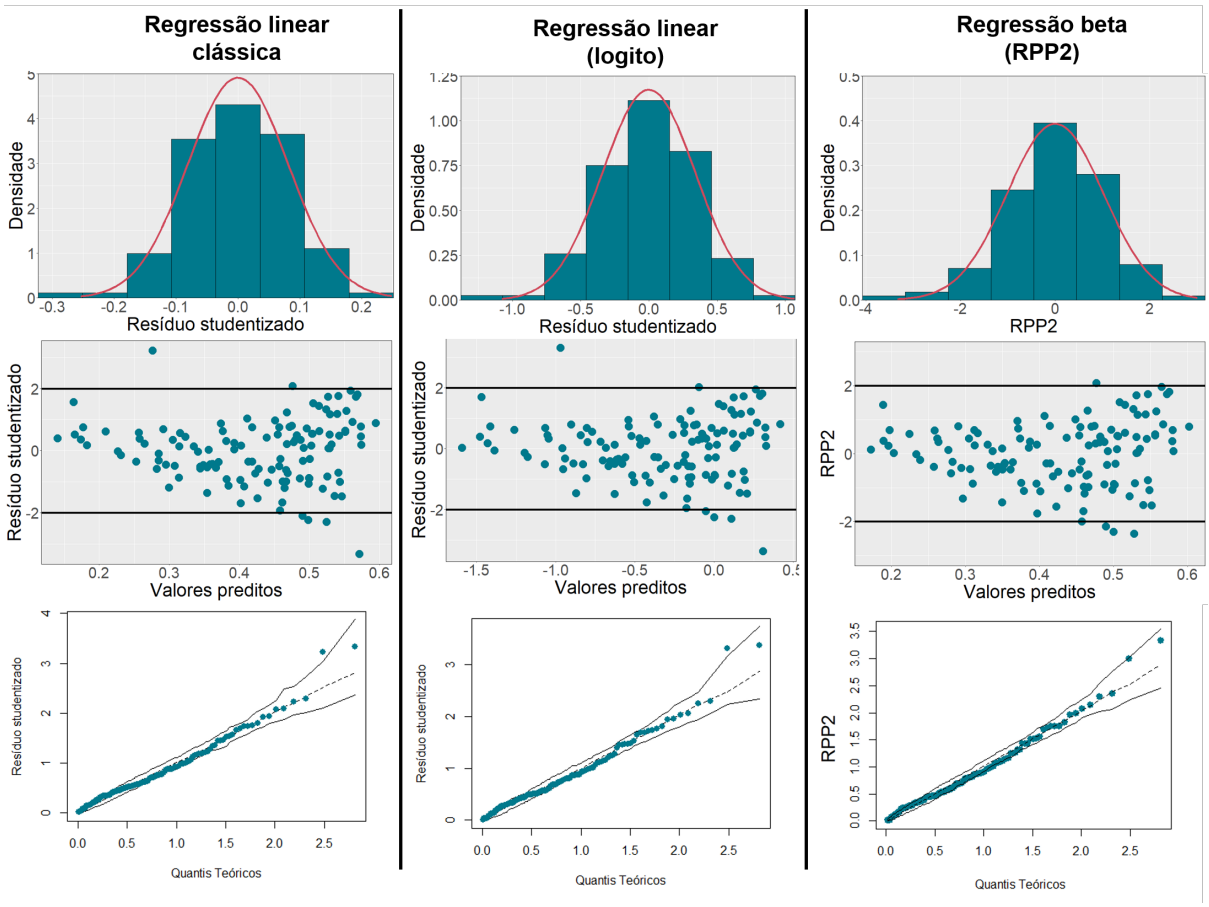

Figura 5.5: Análise de resíduos dos modelos globais.

Ainda sobre os resíduos dos modelos globais gerados, é importante observar como estes se comportam ao longo do espaço, buscando identificar a componente de variação espacial. Por conta disso, a Figura 5.6 apresenta o Mapa de Moran e o Mapa LISA, que indicam a presença de dependência espacial nos resíduos de todos os modelos desenvolvidos.

Vale ressaltar que essa visualização auxilia em identificar locais que tem resíduos altos (ou baixos) e tem vizinhos com resíduos também altos (ou baixos).

Além da análise visual, calculando o I de Moran para os resíduos dos três modelos desenvolvidos, tem-se os valores apresentados na Tabela 5.6, que apontam a dependência espacial em todos os modelos. Por conta disso, fica ainda mais eminente a necessidade de um modelo que leve em consideração o fator espacial, como apresentado na seção seguinte.

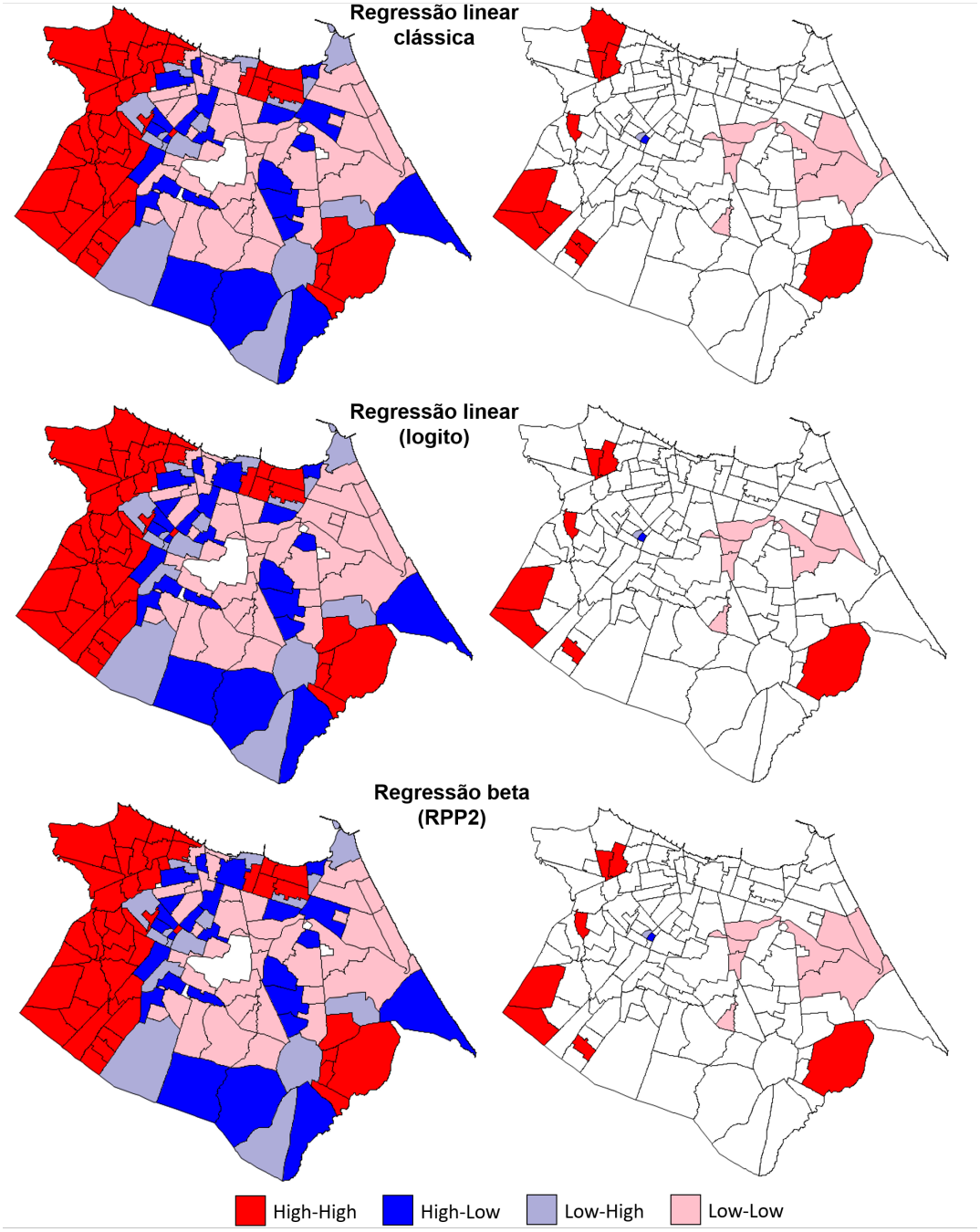

Figura 5.6: Mapa de Moran e Mapa LISA relativos aos resíduos dos modelos globais.

Tabela 5.6: I de Moran para os resíduos dos modelos globais.

Modelo	I de Moran	p-valor
Regressão linear clássica	$0,2594$	$< 0,0001$
Regressão linear (logito)	$0,2330$	$< 0,0001$
Regressão beta (RPP2)	$0,2354$	$< 0,0001$

5.5 Ajuste de modelos locais

Verificada a necessidade da modelagem com a adição da localização como um fator, foram desenvolvidos modelos geograficamente ponderados considerando a distribuição normal sem transformação na variável resposta, nesta Seção tratado apenas pela sigla RGP, um modelo ainda com distribuição normal porém agora transformado pela função logito, denominado aqui por RGP-logito e também a distribuição beta com a função de ligação logito (RBGP), como comentado anteriormente nos métodos do trabalho.

É importante saber que a maior distância possível entre duas Zonas de Tráfego é de 21,62 km, e o número máximo de vizinhos em um modelo é de 125, dado que existem 126 Zonas na cidade de Fortaleza. Um parâmetro de suavização com valor muito próximo à distância máxima entre localidades (ou ao número total de localidades, no caso de uma abordagem adaptável) indica um modelo igual à regressão global, realizada na seção anterior, uma vez que o modelo local consideraria todas as localidades como influências.

5.5.1 Modelo RGP

Inicialmente, utilizando o algoritmo GSS, foram obtidos os parâmetros de suavização apresentados na Tabela 5.7 para o modelo geograficamente ponderado (RGP) normal.

Enquanto que para o método fixo foram obtidos os valores, em metros, 610,29 e 2125,06, otimizando o AIC e o CV, respectivamente, para o método adptável foram obtidos os parâmetros 7 a partir do AIC e 31 pelo CV, que representam o número de vizinhos considerados em cada modelo. O melhor valor a ser usado pode ser identificado apenas quando executados os

modelos de regressão geograficamente ponderada considerando a distribuição normal com cada um desses parâmetros.

Tabela 5.7: Ajuste de modelos RGP para diferentes parâmetros de suavização.

Métrica estatística	**GSS Fixo**		**GSS Adptável**	
	AIC	CV	AIC	CV
Parâmetro de suavização	$610,2894$	$2125,0616$	7	31
R^2 ajustado	$0,8807$	$0,8079$	$0,8778$	$0,8144$
RMSE	$0,0471$	$0,0598$	$0,047613$	$0,0587$
N° de parâmetros efetivos	$120,3938$	$42,6276$	$118,0247$	$42,9665$
Log-verossimilhança	$402,3956$	$202,1134$	$378,7150$	$204,7732$
AIC	$-564,0036$	$-318,9716$	$-521,3807$	$-323,6134$
AICc	$5781,7752$	$-273,8171$	$3506,4803$	$-277,5568$

Em vermelho, o melhor valor encontrado para a estatística.

Comparando as métricas de qualidade do modelo quando ajustados com os quatro diferentes parâmetros de suavização, tem-se a Tabela 5.7.

Quando observado um menor número de parâmetros efetivos, o melhor modelo é o obtido com o GSS fixo minimizando o CV porém, essas estatísticas estão muito próximas às obtidas com o GSS adptável, que tem como vantagem um menor AICc e um maior R^2. Portanto, o modelo selecionado é aquele que considera uma vizinhança de 31 locais, valor obtido com o GSS adptável pela minimização do CV.

Se observado apenas o R^2 ajustado, o melhor modelo seria o realizado considerando um raio de 610,29 metros porém, esse valor gera um número muito grande de parâmetros efetivos, o que pode acarretar num sobreajuste, também conhecido pelo termo em inglês "*overfitting*", que indica um modelo razoável apenas para aquele conjunto de dados, se mostrando ineficaz para a previsão de novas ocorrências.

Como no modelo local não se tem apenas uma estimativa para cada variável, e sim, estimativas para cada Zona de Tráfego estudada, a Tabela 5.8 apresenta algumas estatísticas descritivas dos parâmetros do modelo.

Essas medidas mostram a variabilidade do modelo de cada local, indicando que as variáveis

Tabela 5.8: Resultados do modelo RGP.

Variáveis / Estatísticas	**Estimativas**					
	Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
Intercepto	$0,0813$	$0,1787$	$0,2213$	$0,2455$	$0,2730$	$0,5628$
P_D_A3SM	$-0,1066$	$0,1997$	$0,3128$	$0,3171$	$0,4661$	$0,6651$
DEN_I_SEM	$-0,3807$	$-0,1347$	$-0,0738$	$-0,0945$	$-0,0257$	$0,0252$
D_EQUI_FE	$-1,1821$	$-0,2116$	$-0,0899$	$-0,0755$	$0,0835$	$0,6184$
R^2 ajustado	$0,8144$					
Log-verossimilhança	$204,7732$					
AICc	$-277,5568$					

agem de forma diferente em cada local no espaço. Para as três variáveis explicativas utilizadas no modelo, em alguns locais tem-se uma estimativa de coeficiente negativo enquanto que em outros, o mesmo coeficiente tem um sinal positivo, mudando assim a interpretação do problema como um todo.

Contudo, é necessário investigar primordialmente a significância dessas estimativas, evitando utilizar componentes não significativos para o parâmetro estimado naquela Zona de Tráfego. O comportamento das estimativas significativas no espaço, considerando 10% de significância global, que indica uma significância de 0,94% para cada local $\left(\frac{0,10}{(42,97/4)} = 0,0093\right)$, pode ser visto na Figura 5.7.

A proporção de domicílios com renda familiar abaixo de três salários mínimos possui significância em quase toda a área entretanto, mesmo que significante em toda a área, os parâmetros obtidos são diferentes entre locais, variando de 0,194 a 0,665.

Isso significa dizer que, mantendo todas as outras variáveis explicativas constantes, para um determinado local que fica mais ao nordeste da cidade (como visto na Figura 5.7, a cada acréscimo de 10% na proporção de domicílios de baixa renda, a taxa de acidentes com vítimas cresce cerca de 1,9 pontos percentuais. Já para locais mais periféricos de Fortaleza, esse crescimento pode chegar a 6,6 pontos percentuais.

Já o número de interseções semaforizadas e o número de equipamentos de fiscalização, ambos por km de via, possuem pequenas áreas de significância, indicando assim que, esse fator

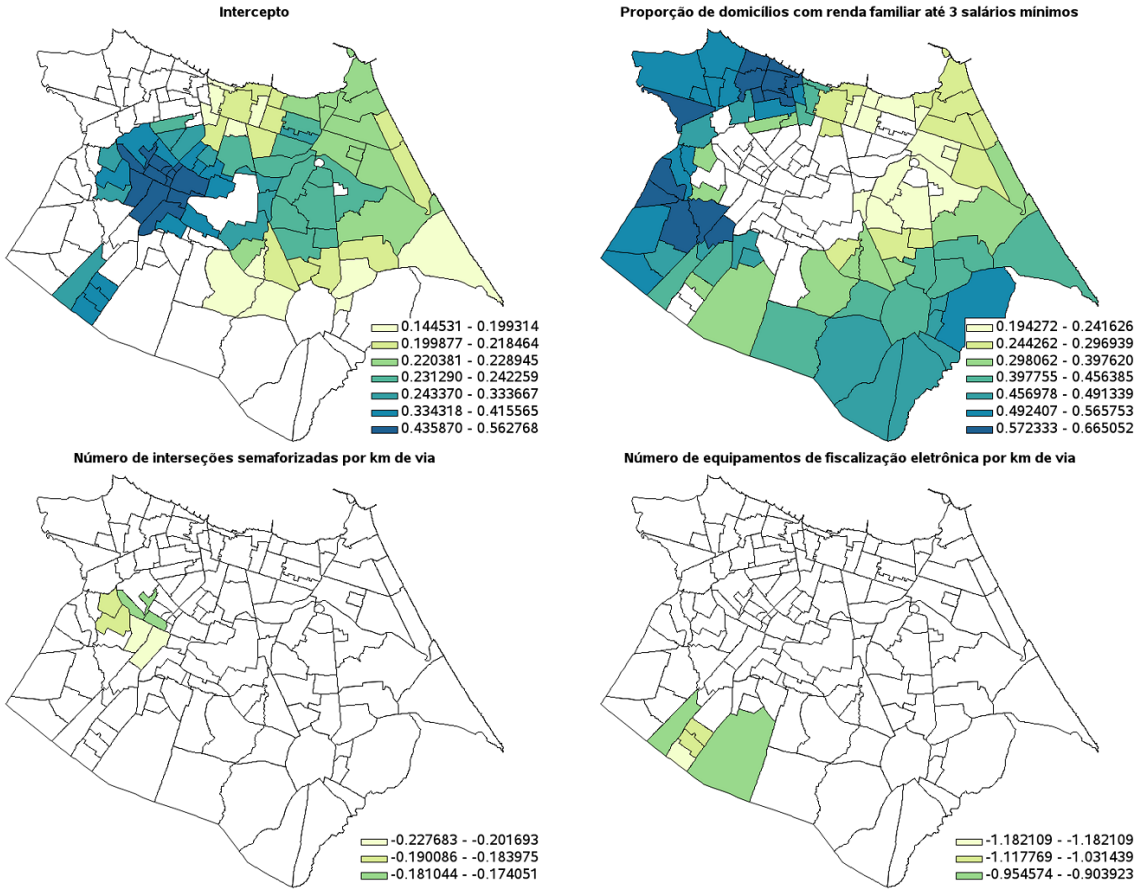

Figura 5.7: Locais com estimativas significantes para cada variável no modelo RGP.

não é necessariamente importante na predição da taxa de acidentes de trânsito com vítimas em todos os locais, apenas em uma pequena delimitação.

Caso seja acrescido uma interseção semaforizada por km de via, espera-se que, em determinados locais, a taxa de acidentes com vítimas caia cerca de 22,8 pontos percentuais e em outros essa mesma taxa caia apenas 17,4 pontos percentuais.

Se considerada a adição de um equipamento de fiscalização eletrônica por km de via, apenas cinco ZT's possuem estimativas significantes, com valores entre -0,95 e -1,18, ou seja, espera-se um decréscimo de 0,95 a 1,18 na taxa de acidentes, a depender da localidade.

É interessante notar que na área onde o intercepto tem maiores valores, a proporção de domicílios com baixa renda não tem significância porém, com a adição de equipamentos de fis-

calização e de semáforos nas interseções, espera-se que essa alta taxa de acidentes com vítimas decaia.

Após ajustar os modelos em cada localidade, os resíduos devem ser analisados, verificando principalmente a ausência da dependência espacial.

Visualmente, pode-se observar a Figura 5.8, que indica em seu primeiro mapa um espalhamento dos valores *low-low*, *high-high*, *low-high* e *high-low*, sem padrão aparente, dividindo bem a cidade estudada entre as categorias.

No mapa LISA, que busca os locais com influência de vizinhança, e com isso dependência espacial, significativa, nota-se que apenas sete ZT's possuem influência significativa dos vizinhos, representando assim apenas 5,5% das localidades em estudo.

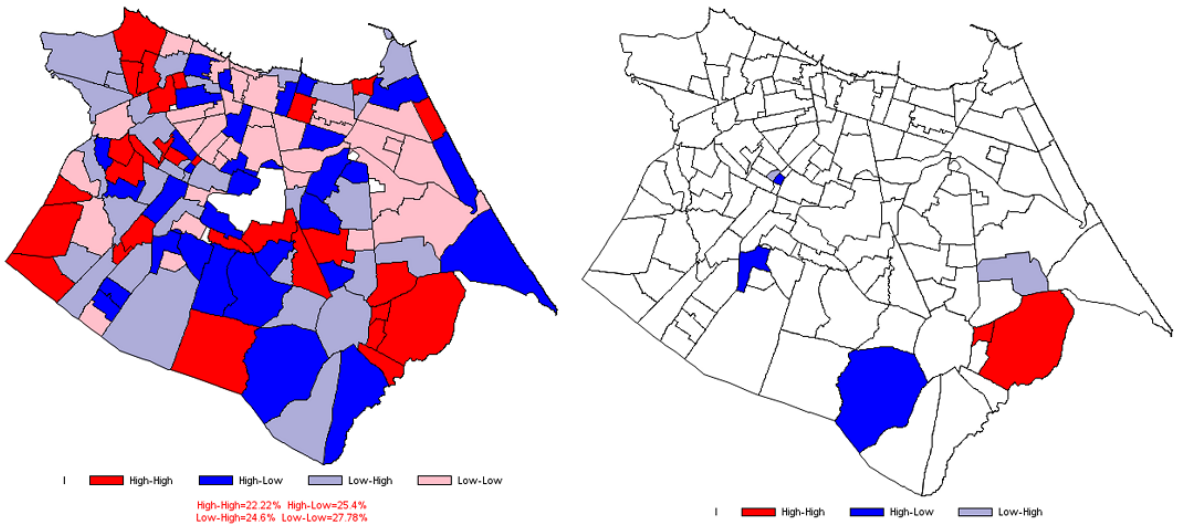

Figura 5.8: Mapa de Moran e Mapa LISA relativos aos resíduos do modelo RGP.

Numericamente, pode ser calculado o I de Moran para os resíduos juntamente com seu p-valor, como apresentado na Tabela 5.9. Com isto, obtém-se um $I = -0,050045$ para os resíduos do modelo gaussiano local e, realizando um teste de hipóteses, não existem evidências que indiquem que os dados possuem correlação espacial, usando uma significância de 10%.

Tabela 5.9: I de Moran para os resíduos do modelo RGP.

Modelo	**I de Moran**	**p-valor**
RGP	$-0,0500$	$0,1506$

5.5.2 Modelo RGP-logito

Agora, buscando aplicar o modelo de regressão geograficamente ponderada com a variável resposta transformada pela função de ligação logito (RGP-logito), foi utilizado o algoritmo GSS, assim como anteriormente. Os parâmetros de suavização e as métricas geradas pelos modelos com cada um desses parâmetros são apresentados na Tabela 5.10.

Tabela 5.10: Ajuste de modelos RGP-logito para diferentes parâmetros de suavização.

Métrica estatística	GSS Fixo		GSS Adaptável	
	AIC	CV	AIC	CV
Parâmetro de suavização	$610,2894$	$2233,692$	7	31
R^2 ajustado	$0,8780$	$0,8037$	$0,8778$	$0,8154$
RMSE	$0,2051$	$0,2602$	$0,2053$	$0,2523$
Nº de parâmetros efetivos	$120,3938$	$39,7204$	$118,0247$	$42,9665$
Log-verossimilhança	$216,9325$	$14,7101$	$194,5948$	$21,0155$
AIC	$-193,0776$	$50,0206$	$-153,1402$	$43,9022$
AICc	$6152,7012$	$87,9531$	$3874,7207$	$89,9587$

Em vermelho, o melhor valor encontrado para a estatística.

Os parâmetros de suavização obtidos são bem próximos dos apresentados na Tabela 5.7 para o modelo RGP sem modificação na variável resposta porém, agora o melhor modelo a ser escolhido, por possuir melhor métrica de AICc e não elevar consideravelmente o número de parâmetros efetivos a serem estimados, é o obtido a partir da minimização do CV com parâmetro fixo, indicando que cada modelo local contará com os vizinhos dentro de um raio de 2,23 km.

As medidas descritivas para o então modelo selecionado são demonstradas a seguir, na Tabela 5.11.

Assim como no modelo RGP ajustado anteriormente, o modelo RGP-logito possui uma grande variabilidade nos parâmetros estimados localmente, carregando diferentes interpretações quanto à interferência da variável explicativa na taxa de acidentes com vítimas. Porém, algumas

Tabela 5.11: Resultados do modelo RGP-logito.

Variáveis/ Estatísticas	**Estimativas**					
	Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
Intercept	$-2,1344$	$-1,2808$	$-1,1822$	$-1,1061$	$-0,9099$	$-0,0545$
P_D_A3SM	$0,6508$	$1,0766$	$1,3562$	$1,4597$	$1,7761$	$2,4833$
DEN_I_SEM	$-1,9619$	$-0,5759$	$-0,3657$	$-0,4262$	$-0,1824$	$0,4713$
D_EQUI_FE	$-6,7153$	$-0,6122$	$-0,4165$	$-0,3657$	$0,2677$	$7,8928$
R^2 ajustado	$0,8037$					
Log-verossimilhança	$14,7101$					
AICc	$87,9531$					

dessas estimativas que geram as medidas apresentadas na Tabela 5.11 não são significativas para os modelos locais e portanto, não devem ser utilizadas na análise. Com isso posto, são gerados os mapas apresentados na Figura 5.9, apresentando somente os coeficientes significativos à 10% de significância global, ou seja, 1,01% de significância local $\left(\frac{0,10}{(39,72/4)} = 0,0101\right)$, e como eles se distribuem no espaço estudado.

Vale ressaltar que a partir dessa transformação, tem-se os parâmetros estimados para o $\log(y_t/(1-y_t))$ e não mais para o y_t.

Assim como no modelo RGP, a variável que é significativa em mais locais é a proporção de domicílios com renda familiar até 3 salários mínimos, com os coeficientes significativos estimados variando de 0,75 a 2,47. A variação dos parâmetros significativos estimados para o número de interseções semaforizadas por km de via é menor, indo de -0,63 a -0,39.

Por fim, analisando o número de equipamentos de fiscalização eletrônica por km de via, variável que possui menos locais significantes do que as outras, tem-se uma variação no parâmetro estimado entre -6,56 e -4,47. O efeito dessa variável nas regiões em que ela é significativa aparenta ter uma mudança ínfima, o que indicaria a não necessidade de um modelo local porém, como dito anteriormente, esse efeito só é válido para uma pequena região de Fortaleza. Usando um modelo global, seria considerado um efeito a nível geral, para toda a cidade, o que não é verdade.

É curioso observar ainda que para essa última variável as localidades com parâmetros esti-

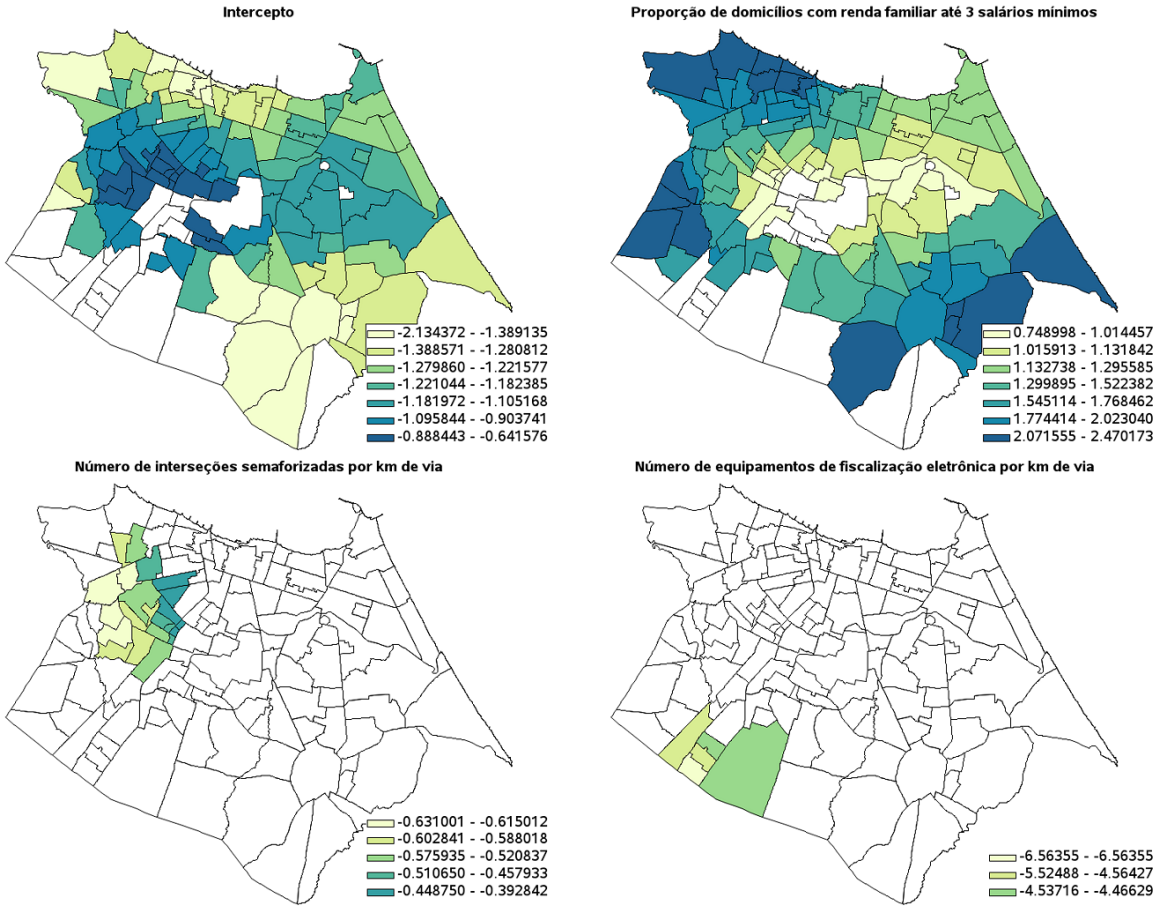

Figura 5.9: Locais com estimativas significantes para cada variável no modelo RGP-logito.

mados significativos não possuem outros parâmetros significantes para nenhuma outra variável, nem mesmo o intercepto, podendo assim então ser essa um local de difícil explicabilidade de ocorrência de acidentes com vítimas a partir das variáveis selecionadas.

Analisando os resíduos distribuídos no espaço a partir do mapa apresentado na Figura 5.10 nota-se a ausência de dependência espacial, fator então captado e explicado pelo modelo. A Tabela 5.12 confirma essa ideia visual a partir da não-rejeição da hipótese nula de independência espacial entre os resíduos, a partir da obtenção de um I de Moran de -0,054277.

Tabela 5.12: I de Moran para os resíduos do modelo RGP-logito.

Modelo	**I de Moran**	**p-valor**
RGP-logito	$-0,0543$	$0,1337$

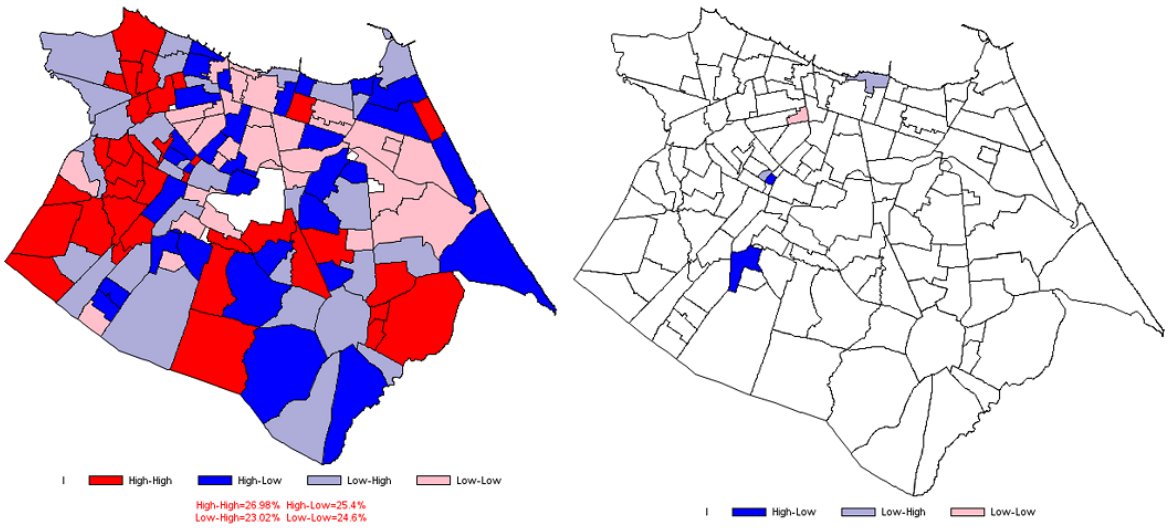

Figura 5.10: Mapa de Moran e Mapa LISA relativos aos resíduos do modelo RGP-logito.

5.5.3 Modelo RBGP

Ajustando, por fim, a regressão beta geograficamente ponderada, foi inicialmente utilizado o algoritmo GSS, obtendo os seguintes parâmetros de suavização, juntamente com as métricas utilizando cada um desses parâmetros (Tabela 5.13).

Tabela 5.13: Ajuste de modelo beta geograficamente ponderado.

Métrica estatística	**GSS Fixo**		**GSS Adaptável**	
	AIC	CV	AIC	CV
Parâmetro de suavização	$11873,6420$	$5103,3176$	125	118
Pseudo R^2 ajustado	$0,6750$	$0,7345$	$0,6357$	$0,6661$
N° de parâmetros efetivos	$4,8197$	$8,9650$	$3,9794$	$4,0999$
Log-verossimilhança	$147,9140$	$164,2649$	$141,4815$	$140,5434$
AICc	$-284,2657$	$-301,7104$	$-278,9719$	$-278,6802$

Em vermelho, o melhor valor encontrado para a estatística.

Selecionando o modelo com os melhores pseudo R^2 ajustado, log-verossimilhança e AICc, tem-se um parâmetro de suavização de aproximadamente 5,1 km, sendo este então, o raio utilizado a partir de cada local para a construção de cada modelo.

Algo a se reparar é que, nos valores propostos pelo GSS adaptável, tem-se modelos muito próximos ao global, considerando quase todos os locais para a predição daquela Zona de Trá-

fego selecionada. Nesses casos, o número de parâmetros efetivos é bem baixo, próximo a 4, representado os parâmetros do modelo global.

Com o modelo desenvolvido a partir do parâmetro de suavização obtido pelo GSS fixo com minimização do CV, tem-se as medidas descritivas apresentadas na Tabela 5.14 para o modelo RBGP.

Tabela 5.14: Resultados do modelo RBGP.

Variáveis/ Estatísticas	**Estimativas**					
	Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
Intercepto	$-1,3149$	$-1,2626$	$-1,2429$	$-1,2434$	$-1,2264$	$-1,1576$
P_D_A3SM	$1,3813$	$1,5769$	$1,7239$	$1,7031$	$1,8207$	$1,9943$
DEN_I_SEM	$-0,5178$	$-0,3495$	$-0,2752$	$-0,2807$	$-0,2099$	$-0,1188$
D_EQUI_FE	$-1,4966$	$-0,3725$	$-0,2537$	$-0,3296$	$-0,1812$	$0,1280$
ϕ	$34,5421$	$37,6043$	$40,8687$	$41,9113$	$43,3702$	$64,7619$
Pseudo R^2 ajustado	$0,7345$					
Log-verossimilhança	$164,2650$					
AICc	$-301,7104$					

Nota-se uma menor variação nos parâmetros estimados quando comparado o modelo RBGP aos modelos locais anteriormente apresentados. Enquanto que para o modelo RGP-logito apresentado na Seção 5.5.2, a Tabela 5.11 mostra o intercepto estimado variando entre -2,13 e -0,05, para o modelo RBGP essa estimativa varia apenas de -1,31 a -1,16. A mesma diminuição de intervalo ocorre para as outras variáveis estudadas.

Um outro parâmetro adicionado ao modelo RBGP é o ϕ, que representa um controle na precisão da distribuição estimada. Considerando um α fixo, quanto maior o valor de ϕ, menor a variância da distribuição, como observado na Figura 3.2.

A interpretação dos parâmetros ocorre de forma similar à feita no modelo RGP-logito, utilizando a razão de chances para identificar o efeito de um fator na variável resposta. A fim de considerar apenas os parâmetros estimados significativos para a análise, com significância local de 5,58% $\left(\frac{0,10}{(8,965/5)} = 0,0558\right)$, tem-se a Figura 5.11, que demarca apenas essas regiões.

Para essa Figura, foram agregados os mapas que representam os parâmetros significativos

para as variáveis “Número de interseções semaforizadas por km de via” e “Número de equipamentos de fiscalização eletrônica por km de via”, que não possuem coloração pois não foram encontradas regiões onde essas variáveis foram significativas.

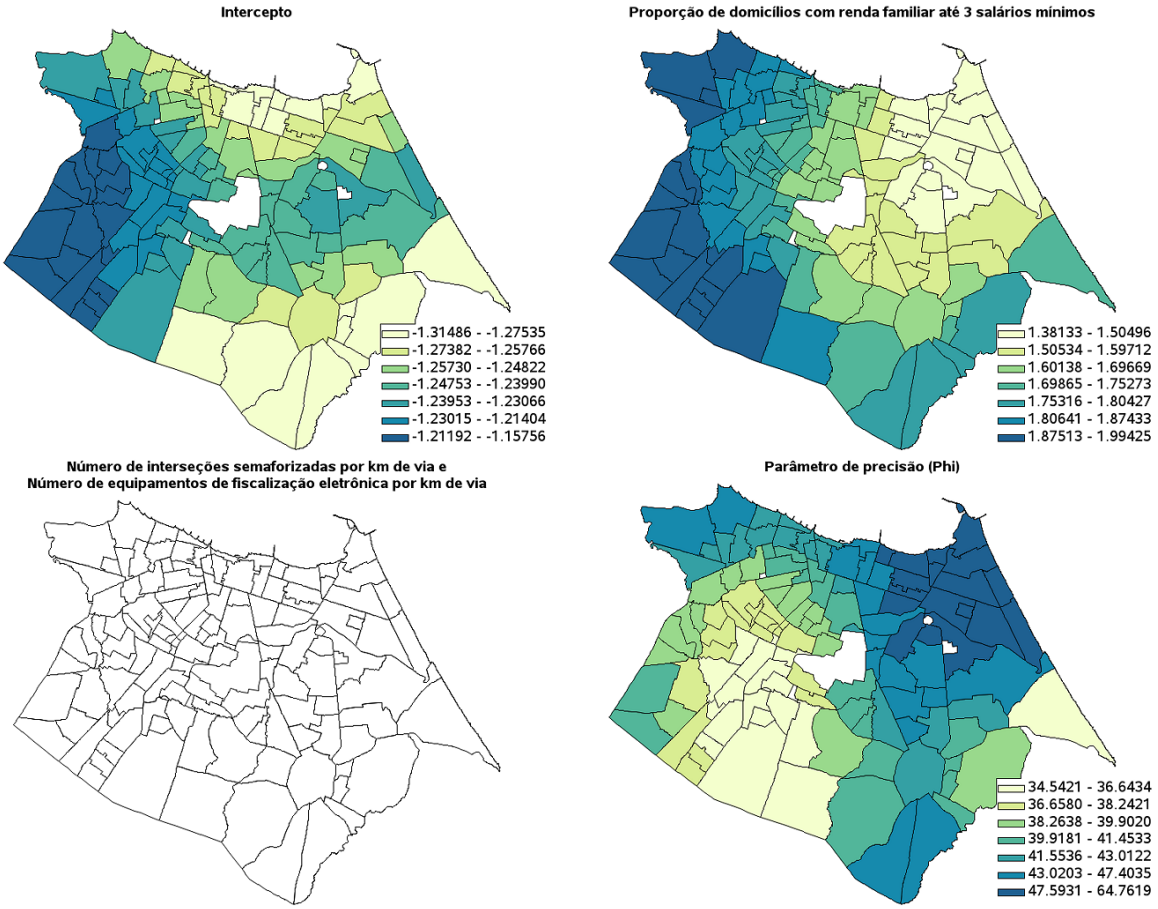

Figura 5.11: Locais com estimativas significantes para cada variável no modelo RBGP.

O modelo RBGP traz o intercepto, a proporção de domicílios de baixa renda e o parâmetro de precisão globalmente significativos porém, com valores estimados diferente para cada local.

No intercepto, existe uma variação entre -1,31 e -1,15, significando que a chance da existência de vítimas em um acidente, sem considerar variáveis explicativas, varia entre 26,9% ($e^{-1,31}$) e 31,4% ($e^{-1,15}$), a depender do local analisado. Com o primeiro mapa da Figura 5.11 é possível perceber que em média os valores mais altos ocorrem mais a oeste do município de Fortaleza, indicando assim que, sem a interferência de outras variáveis explicativas, esse é o local mais

costumeiro para a ocorrência de acidentes com vítimas.

Quando adicionado ao modelo a informação da proporção de domicílios com renda familiar de até 3 salários mínimos, que tem parâmetros estimados entre 1,38 e 1,99, espera-se então que, com o crescimento em 10 pontos percentuais na proporção de domicílio de menor renda, para alguns lugares a taxa de acidentes com vítimas cresça 14,8% ($e^{1,38133\times0,1} - 1$) e, para outros lugares, esse crescimento possa chegar a 22,1% ($e^{1,99425\times0,1} - 1$).

Como dito anteriormente, o número de interseções semaforizadas e o número de equipamentos de fiscalização eletrônica por km de via não são informações significativas em nenhuma localidade. Isso significa dizer que, para o modelo RBGP, essas informações não são relevantes para entender a taxa de acidentes com vítimas na cidade de Fortaleza como um todo.

O parâmetro de precisão varia entre 34,5 e 64,8, valores consideravelmente altos, que indicam uma baixa variabilidade nas estimativas obtidas. Esses valores também estão organizados por local, onde se percebe uma concentração de valores mais altos na região leste e central de Fortaleza, com esses valores decaindo conforme se afasta do centro e se aproxima do oeste da cidade.

Analisando os resíduos ponderados padronizados 2, conforme pode ser visto com os dados apresentados na Tabela 5.15, a independência espacial é obtida considerando 3% de significância. Mesmo com uma grande redução do I de Moran quando comparado ao obtido no modelo global (Tabela 5.6), nota-se que ainda existe um resquício de dependência espacial nos resíduos. Essa característica foi também observada por Da Silva e Lima (2017).

Tabela 5.15: I de Moran para os resíduos do modelo RBGP.

Modelo	**I de Moran**	**p-valor**
RBGP (RPP2)	$0,1113$	$0,0329$

Esse resíduo está distribuído no espaço conforme Figura 5.12, que demonstra alguns poucos locais ainda com a taxa de acidentes com vítimas influenciada pelo seus vizinhos.

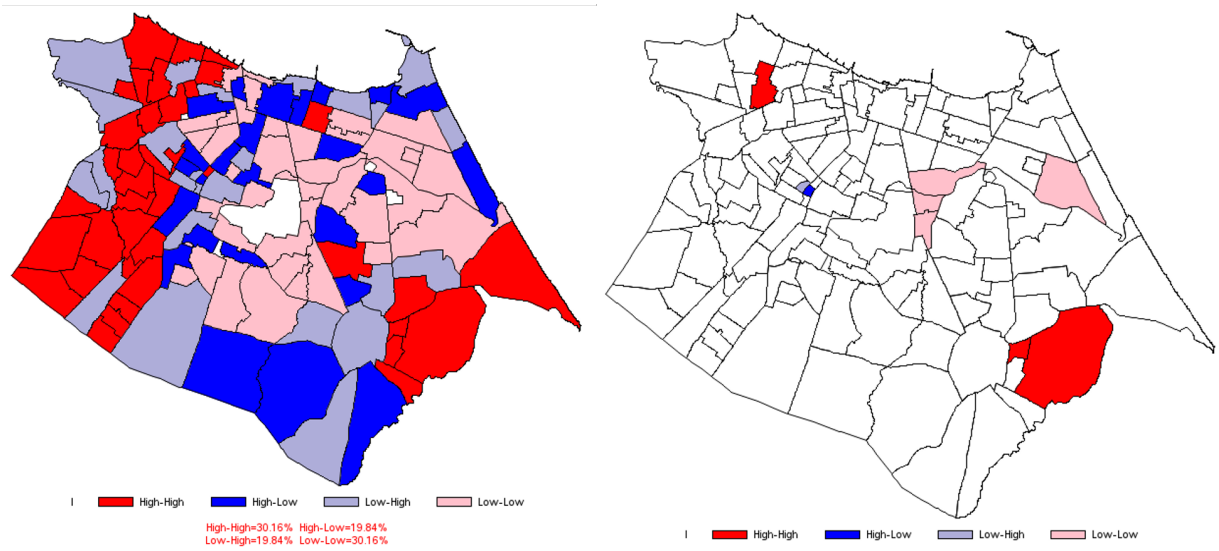

Figura 5.12: Mapa de Moran e Mapa LISA relativos aos RRP2 do modelo RBGP.

5.6 Comparação entre modelos

Analisando as estatísticas obtidas para todos os modelos desenvolvidos até aqui, sejam os globais ou os locais, tem-se a Tabela 5.16.

Tabela 5.16: Métricas de qualidade dos modelos ajustados.

Métrica estatística	**Modelo**					
	Global			**Local**		
	Normal	Normal-logito	Beta	RGP	RGP-logito	RBGP
R^2 ajustado*	$0,6357$	$0,6563$	$0,6535$	$0,8144$	$0,8037$	$0,7345$
Nº de parâmetros efetivos	-	-	-	$42,9665$	$39,7204$	$12,0237$
Log-verossimilhança	$138,0389$	$-42,3817$	$143,5261$	$204,7732$	$14,7101$	$164,2649$
AIC	$-268,0778$	$92,7634$	$-277,0700$	$-323,6134$	$50,0206$	$-304,4825$
AICc	$-267,7472$	$93,0940$	$-276,5724$	$-277,5568$	$87,9531$	$-301,7104$

Em vermelho, o melhor valor encontrado para a estatística.
*Para o modelo RBGP, pseudo R^2 ajustado.

Dentre os modelos aplicados, o que mostra melhor R^2 ajustado foi o modelo de regressão geograficamente ponderada (RGP) porém, essas métricas podem trazer algumas conclusões imprecisas, uma vez que o R^2 ajustado para o modelo RBGP, usado na comparação, é um pseudo R^2, usado apenas para uma idealização da métrica de explicação e não para um resultado objetivo.

Além disso, quando observado o número de parâmetros estimados, nota-se um número 3,6 vezes maior para o modelo RGP se comparado ao modelo de regressão beta geograficamente ponderado (RBGP). Isso pode indicar a ocorrência de *overfitting*, fazendo com que esse modelo sirva apenas para os dados apresentados, sendo impreciso para a estimação de novos casos. Para confirmar essa hipótese, poderia ser observado um gráfico de envelope simulado, ainda não desenvolvido para modelos locais.

É curioso ainda observar que os modelos que não estimam o parâmetro de dispersão (ϕ) junto a ele possuem mais parâmetros estimados do que o modelo RBGP, que estima o parâmetro ϕ. Esse apontamento também é visualizado em Da Silva e Rodrigues (2014), com o uso da distribuição binomial negativa.

Sobretudo, vale ressaltar que a comparação entre modelos por meio da log-verossimilhança, do AIC e do AICc não pode ser realizada, uma vez que os modelos são construídos a partir de diferentes técnicas.

Um outro fator importante para o processo é a interpretabilidade dos parâmetros. Entre os modelos locais, apenas o RBGP têm seus parâmetros como razões de chance, podendo assim obter a chance de aumento de ocorrências em função da comparação entre diferentes panoramas. Essa informação é ainda mais essencial no modelo local, onde tem-se distintos acréscimos (ou decréscimos) de chance para cada região analisada, podendo assim então saber o efeito esperado de determinada ação em cada localidade.

Caso o modelo seja usado apenas para predição esse fator não é importante porém, quando se deseja entender os motivos da ocorrência de eventos, como aqui neste trabalho, é importante utilizar um modelo que seja de fácil interpretação. Um exemplo disso é o modelo RGP-logito, que mesmo tendo um bom valor de R^2, traz informações apenas sobre o $\log(y/(1-y))$ dificultando assim as interpretações diretas à taxa de acidentes com vítimas.

A Tabela 5.17 apresenta as informações relativas à independência espacial do processo. É eminente a necessidade do uso de um modelo local, visto que a hipótese de independência espacial dos resíduos é rejeitada em todos os modelos globais.

Tabela 5.17: Comparação da independência espacial dos resíduos dos modelos desenvolvidos.

Tipo	**Modelo**	**I de Moran**	**p-valor**
Global	Regressão linear clássica	$0,2594$	$< 0,0001$
	Regressão linear (logito)	$0,2330$	$< 0,0001$
	Regressão beta (RPP2)	$0,2354$	$< 0,0001$
Local	RGP	$-0,0500$	$0,1506$
	RGP-logito	$-0,0543$	$0,1337$
	RBGP (RPP2)	$0,1113$	$0,0329$

Em linhas amarelas, modelos com dependência espacial a qualquer nível de significância.

Nos modelos locais é possível anular o componente de espaço quando usada uma significância de cerca de 3%. Nota-se que o I de Moran obtido para o modelo RBGP, por exemplo, é 2,1 vezes menor do que o obtido no modelo de regressão beta global, mostrando mais uma vez a necessidade de um modelo local para a análise do caso apresentado.

Capítulo 6

Conclusões

A principal motivação deste trabalho foi trazer a aplicação de uma modelagem de dados apresentados como uma taxa juntamente com uma abordagem local, que possibilita considerar e entender o processo de diferentes formas dentre o espaço.

Uma vez que o modelo local retorna ao modelo global na ausência de qualquer dependência espacial, conclui-se que os modelos locais são sempre mais informativos ou tão informativos quanto os modelos globais. Além disso, ganha-se muito na interpretabilidade do modelo de acordo com a localidade, e ainda obtêm-se vantagem na visualização dos resultados, uma vez que agora, com o modelo local, é possível usar o recurso cartográfico.

Para o estudo de caso, foram considerados modelos globais e locais, confirmando a hipótese de que os modelos locais seriam mais informativos e teriam uma melhor performance. Quanto à distribuição, para a taxa de acidentes com vítimas, que tem um desenho simétrico, não houve ganho significativo no uso da distribuição beta, a não ser pela interpretação dos parâmetros via razão de chance.

Contudo, a abordagem com o uso da distribuição beta garante a adequabilidade do modelo tanto à distribuições assimétricas, como visto em Da Silva e Lima (2017), quanto à distribuições simétricas, como aqui foi apresentado. Por conta disso, se tratando da análise de taxas, é sempre recomendado o uso da distribuição beta, uma vez que ela se adéqua às mais diversas formas de

distribuição que podem ser apresentadas.

Além dos resultados obtidos com a aplicação do modelo RBGP foi desenvolvido um pacote no *software* R denominado *"gwbr"*, com o intuito de difundir o uso da técnica aqui aplicada, uma vez que a ferramenta é aberta e pode ser usada de forma gratuita. O pacote está disponibilizado na plataforma de hospedagem, gestão e compartilhamento de códigos-fonte GitHub por meio do caminho `https://github.com/romarq23/gwbr` e também já está disponível no repositório CRAN, podendo assim ser acessado de forma nativa no R. As informações quanto ao pacote podem ser visualizadas em `https://cran.r-project.org/web/packages/gwbr/index.html`.

6.1 Limitações do trabalho

Uma das limitações do trabalho aqui desenvolvido foi a ausência de dados para o objetivo idealizado a princípio. Inicialmente, era desejado trabalhar com dados relativos à mortes em acidentes de trânsito. Por conta da ausência desse tipo de informação, pôde-se trabalhar apenas com acidentes que não geraram somente danos materiais, mas também algum tipo de dano à integridade física do cidadão no veículo. Pode-se levantar a hipótese de que, por conta disso, a distribuição foi mais simétrica, uma vez que se considerados acidentes fatais, espera-se um número mais baixo de ocorrências em relação aos acidentes gerais.

Em relação ao algoritmo criado, uma limitação é o tempo de execução de algumas funções, principalmente o *GSS*. Além disso, como o algoritmo foi traduzido da linguagem SAS para a linguagem R, algumas operações não foram tão eficentes, do ponto de vista computacional, por conta do código ter sido desenvolvido para a execução em um *software* totalmente diferente do implementado.

6.2 Trabalhos futuros

Ainda existe um extenso campo para a evolução dos modelos de regressão geograficamente ponderados. Seguem a seguir algumas sugestões para trabalhos futuros, tanto em relação ao RBGP quanto ao pacote em R desenvolvido:

- Aplicação em casos de acidentes com vítimas fatais;
- Busca de ajuste para a remoção da dependência espacial completa no modelo de regressão beta geograficamente ponderada;
- Aplicação em taxas com distribuições mais assimétricas e/ou de caudas mais pesadas;
- Desenvolvimento de gráficos de envelope simulado para os modelos locais;
- Aplicação da RBGP em taxas com excessos de zeros, valor não suportado pela distribuição beta, exigindo algum tipo de mistura de distribuições;
- Desenvolvimento de um algoritmo "*stepwise*" para a seleção de variáveis para modelos locais;
- Comparação com modelos que consideram abordagens Bayesianas;
- Desenvolvimento e aplicação de novas medidas para a qualidade do modelo;
- Melhoria no tempo de execução dos algoritmos implementados em R;
- Evoluções na entrada dos parâmetros nas funções desenvolvidas em R, deixando mais próximo ao usual de funções nativas da linguagem.

Anexo A

Tabela A.1 - Matriz de correlação com todas as variáveis disponíveis.

	AC_TX	AREA_KM	POP_TOT	EXT_TOT	DEN_I_SEM	DEN_I_NSEM	D_EQUI_FE
AC_TX	1	0,23254	0,36339	0,23801	−0,58062	0,17261	−0,31849
AREA_KM	0,23254	1	0,62304	0,86028	−0,32460	−0,28211	−0,13880
POP_TOT	0,36339	0,62304	1	0,88010	−0,29019	0,12121	−0,07479
EXT_TOT	0,23801	0,86028	0,88010	1	−0,27415	−0,12838	−0,09902
DEN_I_SEM	−0,58062	−0,32460	−0,29019	−0,27415	1	−0,17010	0,40280
DEN_I_NSEM	0,17261	−0,28211	0,12121	−0,12838	−0,17010	1	−0,18925
D_EQUI_FE	−0,31849	−0,13880	−0,07479	−0,09902	0,40280	−0,18925	1
P_0_17	0,73683	0,38617	0,34909	0,32086	−0,68657	0,06035	−0,34003
P_18_64	−0,69517	−0,29683	−0,32209	−0,25457	0,57295	−0,16861	0,32904
P_M64	−0,63875	−0,41037	−0,31076	−0,33209	0,67896	0,07276	0,28555
P_D_A3SM	0,79562	0,25272	0,34761	0,24388	−0,63074	0,27121	−0,32969
P_D_M3SM	−0,79562	−0,25272	−0,34761	−0,24388	0,63074	−0,27121	0,32969
URES_A	−0,52139	−0,33538	−0,16578	−0,24044	0,57491	−0,06547	0,28525
UCOPS_A	−0,57248	−0,28897	−0,23131	−0,19254	0,86743	−0,19764	0,44822

Em vermelho, correlações não significativas considerando 95% de confiança. (Continua)

Tabela A.2 - (Continuação) Matriz de correlação com todas as variáveis disponíveis.

	P_0_17	P_18_64	P_M64	P_D_A3SM	P_D_M3SM	URES_A	UCOPS_A
AC_TX	0,73683	−0,69517	−0,63875	0,79562	−0,79562	−0,52139	−0,57248
AREA_KM	0,38617	−0,29683	−0,41037	0,25272	−0,25272	−0,33538	−0,28897
POP_TOT	0,34909	−0,32209	−0,31076	0,34761	−0,34761	−0,16578	−0,23131
EXT_TOT	0,32086	−0,25457	−0,33209	0,24388	−0,24388	−0,24044	−0,19254
DEN_I_SEM	−0,68657	0,57295	0,67896	−0,63074	0,63074	0,57491	0,86743
DEN_I_NSEM	0,06035	−0,16861	0,07276	0,27121	−0,27121	−0,06547	−0,19764
D_EQUI_FE	−0,34003	0,32904	0,28555	−0,32969	0,32969	0,28525	0,44822
P_0_17	1	−0,91778	−0,89564	0,89365	−0,89365	−0,72879	−0,66322
P_18_64	−0,91778	1	0,64539	−0,89166	0,89166	0,61848	0,54585
P_M64	−0,89564	0,64539	1	−0,72031	0,72031	0,70919	0,66440
P_D_A3SM	0,89365	−0,89166	−0,72031	1	−1	−0,72441	−0,58301
P_D_M3SM	−0,89365	0,89166	0,72031	−1	1	0,72441	0,58301
URES_A	−0,72879	0,61848	0,70919	−0,72441	0,72441	1	0,57692
UCOPS_A	−0,66322	0,54585	0,66440	−0,58301	0,58301	0,57692	1

Em vermelho, correlações não significativas considerando 95% de confiança.

Referências Bibliográficas

Abdel-Aty, M., Siddiqui, C., Huang, H., e Wang, X. (2011). Integrating trip and roadway characteristics to manage safety in traffic analysis zones. *Transportation Research Record*, 2213(1):20–28.

Abdel-Aty, M. A. e Radwan, A. E. (2000). Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention*, 32(5):633–642.

Aguero-Valverde, J. e Jovanis, P. P. (2006). Spatial analysis of fatal and injury crashes in pennsylvania. *Accident Analysis & Prevention*, 38(3):618–625.

Albuquerque, P. H. M., Medida, F. A. S., e Da Silva, A. R. (2017). Geographically weighted logistic regression applied to credit scoring models. *Revista Contabilidade & Finanças*, 28(73):93–112.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Dordrecht.

Anselin, L. (1995). Local Indicators of Spatial Association - LISA. *Geographical Analysis*, 27(2):93–115.

Anselin, L. (1996). The moran scatterplot as an esda tool to assess local instability in spatial association. *Spatial Analytical Perspectives on GIS*, pages 111–125.

Arditi, D., Lee, D., e Polat, G. (2007). Fatal accidents in nighttime vs. daytime highway construction work zones. *Journal of Safety Research*, 38(4):399–405.

Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika*, 68(1):13–20.

Belin, M., Johansson, R., Lindberg, J., e Tngval, C. (1997). The vision zero and its consequences. Reprint from the proceedings of the 4th international conference on Safety and the Environment in the 21st century, november 23-27, 1997, Tel Aviv, Israel.

Brunsdon, C., Fotheringham, A. S., e Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4):281–298.

Chatterjee, A., Wegmann, F. J., Fortey, N. J., e D., E. J. (2001). Incorporating safety and security issues in urban transportation planning. *Transportation Research Record*, 1777(1):75–83.

Chin, H. C. e Quddus, M. A. (2003). Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. *Accident Analysis & Prevention*, 35(2):253–259.

Cleveland, W. e Devlin, S. (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Conover, W. J. (1980). *Practical Nonparametric Statistics*, (2 ed.). New York: John Wiley & Sons.

Da Silva, A. e Rodrigues, T. (2014). Geographically weighted negative binomial regression - incorporating overdispersion. *Statistics and Computing*, 24(5):769–783.

Da Silva, A. R. e Fotheringham, A. S. (2016). The multiple testing issue in geographically weighted regression. *Geographical Analysis*, 48(3):233–247.

Da Silva, A. R. e Lima, A. O. (2017). Geographically weighted beta regression. *Spatial Statistics*, 21:279–303.

De Andrade, L., Vissoci, J. R. N., Rodrigues, C. G., Finato, K., Carvalho, E., Pietrobon, R., de Souza, E. M., Nihei, O. K., Lynch, C., e Carvalho, M. D. B. (2014). Brazilian road traffic fatalities: A spatial and environmental analysis. *PLOS ONE*, 9(1):1–10.

De Leur, P. e Sayed, T. (2002). Developing systematic framework for proactive road safety planning. Presented at 81st Annual Meeting of the Transportation Research Board, Washington, D.C.

Dunn, P. K. e Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244.

Dyke, G. e Patterson, H. (1952). Analysis of factorial arrangements when the data are proportions. *Biometrics*, 8(1):1–12.

Elderton, W. P. (1906). *Frequency-Curves and Correlation*. Charles and Edwin Layton, London.

Espinheira, P. L., Ferrari, S. L. P., e Cribari-Neto, F. (2008). On beta regression residuals. *Journal of Applied Statistics*, 35(4):407–419.

Ferrari, S. e Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

Figueira, A. C., S., P. C., De Oliveira, P. T. M. S., e Larocca, A. P. C. (2017). Identification of rules induced through decision tree algorithm for detection of traffic accidents with victims: A study case from brazil. *Case Studies on Transport Policy*, 5(2):200–207.

Fletcher, R. e Powell, M. J. D. (1963). A rapidly convergent descent method for minimization. *The Computer Journal*, 6(2):163–168.

Fotheringham, A. S., Charlton, M., e Brunsdon, C. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley.

Fredette, M., Mambu, L. S., Chouinard, A., e Bellavance, F. (2008). Safety impacts due to the incompatibility of suvs, minivans, and pickup trucks in two-vehicle collisions. *Accident Analysis & Prevention*, 40(6):1987–1995.

Gomes, M. J. T. L., Cunto, F., e Da Silva, A. R. (2017). Geographically weighted negative binomial regression applied to zonal level safety performance models. *Accident Analysis & Prevention*, 106:254–261.

Hadayeghi, A., Shalaby, A. S., e Persaud, B. N. (2010). Development of planning level transportation safety tools using geographically weighted poisson regression. *Accident Analysis & Prevention*, 42(2):676–688.

Hadayeghi, A., Shalaby, A. S., Persaud, B. N., e Cheung, C. (2006). Temporal transferability and updating of zonal level accident prediction models. *Accident Analysis & Prevention*, 38(3):579–589.

Hanna, C. L., Laflamme, L., e Bingham, C. R. (2012). Fatal crash involvement of unlicensed young drivers: County level differences according to material deprivation and urbanicity in the united states. *Accident Analysis & Prevention*, 45:291–295.

Huang, H., Abdel-Aty, M. A., e Darwiche, A. L. (2010). County-level crash risk analysis in florida: Bayesian spatial modeling. *Transportation Research Record*, 2148(1):27–37.

Institute for Health Metrics and Evaluation (2019). Global Burden of Disease. Disponível em: `https://ghdx.healthdata.org/gbd-2019`. Acesso em: 29 dez. 2022.

ITE (1982). *Transportation and Traffic Engineering Handbook*. Institute of Transportation Engineers, Second Edition, Prentice Hall, pp. 344–345.

Johansson, R. (2009). Vision zero - Implementing a policy for traffic safety. *Safety Science*, 47(6):826–831.

Jovanis, P. P. e Chang, H. (1986). Modeling the relationship of accident to miles traveled. *Transportation Research Record*, 1068:42–51.

Lambert-Bélanger, A., Dubois, S., Weaver, B., Mullen, N., e Bédard, M. (2012). Aggressive driving behaviour in young drivers (aged 16 through 25) involved in fatal crashes. *Journal of Safety Research*, 43(5):333–338.

Lee, J., Abdel-Aty, M., e Jiang, X. (2014). Development of zone system for macro-level traffic safety analysis. *Journal of Transport Geography*, 38:13–21.

Luenberger, D. G. e Ye, Y. (1984). *Linear and Nonlinear Programming*, (4 ed.). Springer.

Martínez, L. M., Viegas, J. M., e Silva, E. A. (2009). A traffic analysis zone definition: a new methodology and algorithm. *Transportation*, 36:581–599.

Miaou, S. e Lum, H. (1993). Modeling vehicle accidents and highway geometric design relationships. *Accident Analysis & Prevention*, 25(6):689–709.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.

Nakaya, T., Fotheringham, A., Brunsdon, C., e Charlton, M. (2005). Geographically weighted poisson regression for disease association mapping. *Statistics in Medicine*, 24(17):2695–2717.

Neter, J., Wasserman, W., e Kutner, M. H. (1983). *Applied Linear Regression Models*. Richard D. Irwin, Inc. Homewood, Illinois.

O Povo (2021). Fortaleza agora tem 12 Regionais; você sabe qual é a sua? Disponível em: `https://www.opovo.com.br/noticias/fortaleza/2021/01/05/fortaleza-passa-a-ter-12-regionais--voce-sabe-qual-e-a-sua.html`. Acesso em: 12 mai. 2023.

Obelheiro, M. R., Da Silva, A. R., Nodari, C. T., Cybis, H. B. B., e Lindau, L. A. (2020). A new zone system to analyze the spatial relationships between the built environment and traffic safety. *Journal of Transport Geography*, 84:102699.

Openshaw, S. (1977). Optimal zoning systems for spatial interaction models. *Environment and Planning A: Economy and Space*, 9(2):169–184.

Oris, W. N. (2011). Spatial Analysis of Fatal Automobile Crashes in Kentucky. *Masters Theses & Specialist Projects*, Paper 1119.

Pereira, G. H. A. (2019). On quantile residuals in beta regression. *Communications in Statistics - Simulation and Computation*, 48(1):302–316.

Quddus, M. A. (2008). Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of london crash data. *Accident Analysis & Prevention*, 40(4):1486–1497.

Rhee, K.-A., Kim, J.-K., ihn Lee, Y., e Ulfarsson, G. F. (2016). Spatial regression analysis of traffic crashes in seoul. *Accident Analysis & Prevention*, 91:190–199.

Safarpour, H., Khorasani-Zavareh, D., e Mohammadi, R. (2020). The common road safety approaches: A scoping review and thematic analysis. *Chinese Journal of Traumatology*, 23(2):113–121.

Siddiqui, C., Abdel-Aty, M., e Choi, K. (2012). Macroscopic spatial analysis of pedestrian and bicycle crashes. *Accident Analysis & Prevention*, 45:382–391.

Simpson, T. (1740). *Essays on Several Curiousand Useful Subjects in Speculative and Mix'd Mathematicks, Illustrated by a Variety of Examples*. London.

Siskind, V., Steinhardt, D., Sheehan, M., O'Connor, T., e Hanks, H. (2011). Risk factors for fatal crashes in rural australia. *Accident Analysis & Prevention*, 43(3):1082–1088.

Sivak, M., Schoettle, B., e Rupp, J. (2010). Survival in fatal road crashes: Body mass index, gender, and safety belt use. *Traffic Injury Prevention*, 11(1):66–68.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., e Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Tarko, A. P. (2006). Calibration of safety prediction models for planning transportation networks. *Transportation Research Record*, 1950(1):83–91.

Valent, F., Schiava, F., Savonitto, C., Gallo, T., Brusaferro, s., e Barbone, F. (2002). Risk factors for fatal road traffic accidents in udine, italy. *Accident Analysis & Prevention*, 34(1):71–84.

Vision Zero Network (2014). What is Vision Zero? Disponível em: `https://visionzeronetwork.org/about/what-is-vision-zero/`. Acesso em: 11 nov. 2022.

Vision Zero Network (2018). Core Elements for Vision Zero Communities. Disponível em: `https://visionzeronetwork.org/wp-content/uploads/2022/07/Vision_Zero_Core_Elements.pdf`. Acesso em: 11 nov. 2022.

Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics.

World Health Organization (2019). WHO Mortality Database - Road traffic accidents. Disponível em: `https://platform.who.int/mortality/themes/theme-details/topics/indicator-groups/indicator-group-details/MDB/road-traffic-accidents`. Acesso em: 26 dez. 2022.

Xu, P. e H., H. (2015). Modeling crash spatial heterogeneity: Random parameter versus geographically weighting. *Accident Analysis & Prevention*, 75:16–25.

Xu, P., Huang, H., Dong, N., e Abdel-Aty, M. A. (2014). Sensitivity analysis in the context of regional safety modeling: Identifying and assessing the modifiable areal unit problem. *Accident Analysis & Prevention*, 70:110–120.

Zhao, F. e Park, N. (2004). Using geographically weighted regression models to estimate annual average daily traffic. *Transportation Research Record*, 1879(1):99–107.