# ANÁLISE ESPACIAL DE ACIDENTES DE TRÂNSITO NO CONTEXTO DE VARIÁVEIS AGREGADAS EM ÁREAS: PROPOSTA METODOLÓGICA E APLICAÇÃO NA CIDADE DO RIO DE JANEIRO

Marcos de Meneses Rocha

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia de Transportes, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia de Transportes.

Orientador: Carlos David Nassi

Rio de Janeiro

Março de 2015

ANÁLISE ESPACIAL DE ACIDENTES DE TRÂNSITO NO CONTEXTO DE VARIÁVEIS AGREGADAS EM ÁREAS: PROPOSTA METODOLÓGICA E APLICAÇÃO NA CIDADE DO RIO DE JANEIRO

Marcos de Meneses Rocha

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA DE TRANSPORTES.

Examinada por:

_____
Prof. Carlos David Nassi, Dr.Ing.

_____
Prof. Licinio da Silva Portugal, D.Sc.

_____
Profª Marilita Gnecco de Camargo Braga, Ph.D.

_____
Prof. Flávio José Craveiro Cunto, Ph.D.

_____
Prof. Antônio Nelson Rodrigues da Silva, D.Sc.

RIO DE JANEIRO, RJ - BRASIL
MARÇO DE 2015

Rocha, Marcos de Meneses

Análise Espacial de Acidentes de Trânsito no Contexto de Variáveis Agregadas em Áreas: Proposta Metodológica e Aplicação na Cidade do Rio de Janeiro / Marcos de Meneses Rocha. – Rio de Janeiro: UFRJ/COPPE, 2015.

XIII, 156 p.: il.; 29,7 cm.

Orientador: Carlos David Nassi

Tese (doutorado) – UFRJ/ COPPE/ Programa de Engenharia de Transportes, 2015.

Referências Bibliográficas: p. 143-156.

1. Acidentes de trânsito. 2. Análise espacial. 3. Modelos econométricos. I. Nassi, Carlos David. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Transportes. III. Título.

À Deus.

À Daniele Rocha, minha esposa.

À José Maria da Rocha e Célia Rocha, meus pais.

Ao meu país.

**DEDICO.**

## AGRADECIMENTOS

À Deus, por me ter permitido chegar aqui.

Ao Exército Brasileiro, por esta oportunidade ímpar de realização e crescimento profissionais. Aos meus companheiros de trabalho do IME que me motivaram a fazer o doutorado e me apoiaram em todos os momentos do curso.

Ao Programa de Engenharia de Transportes da COPPE/UFRJ, na pessoa de todos os professores e funcionários que aqui me acolheram e muito me enriqueceram com o seu conhecimento, o seu exemplo e seu serviço dedicado. Uma menção especial à Jane e à Helena, que sempre nos acolhem com muita alegria na secretaria e que sempre estão dispostas a nos ajudar sempre que precisamos.

Aos funcionários Fátima, Gilberto e Yuri que muito me apoiaram para que tivesse as melhores condições de trabalho.

Ao Prof. Dr. Ing. Carlos David Nassi pela confiança, orientação segura e apoio. Aprendi muito com você nesses anos. Aprendi como se deve ensinar, orientar e, principalmente, como respeitar e compreender os outros.

Aos companheiros de laboratório, Marcelinho, Fred, Márcia, Sérgio e Cristiano, pelo acolhimento e amizade nestes anos.

Aos meus amigos alunos do PET, Cláudio Falavigna, Vicente Fernandes e Paolo Galli, pela amizade e pelos bate-papos no almoço e cafezinho que me enriqueciam de boas ideias e me ajudavam a refazer as forças para retornar ao trabalho com mais ânimo.

Aos meus pais, pelas orações e por me acompanharem em todos os momentos dessa caminhada. O exemplo de vocês foi sempre um norte para que continuasse a lutar e chegar neste momento.

À minha esposa Daniele, pela paciência e compreensão. Foi você que me motivou a fazer o doutorado quando ainda namorávamos em 2009. E sempre continuou a ser uma grande incentivadora na época de noivado e agora como casados.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

# ANÁLISE ESPACIAL DE ACIDENTES DE TRÂNSITO NO CONTEXTO DE VARIÁVEIS AGREGADAS EM ÁREAS: PROPOSTA METODOLÓGICA E APLICAÇÃO NA CIDADE DO RIO DE JANEIRO

Marcos de Meneses Rocha

Março/2015

Orientador: Carlos David Nassi

Programa: Engenharia de Transportes

Esta tese apresenta uma metodologia para a análise espacial de acidentes de trânsito georreferenciados na cidade do Rio de Janeiro, no contexto de variáveis agregadas em área e que procura contemplar as especificidades dos dados geográficos, tais como dependência espacial, heterogeneidade espacial, a questão do *Modifiable Areal Unit Problem* (MAUP), falácia ecológica e efeito das bordas. É composta por três grandes etapas: aquisição, compreensão da distribuição espacial e modelagem dos dados de acidentes. As variáveis mais explicativas foram a hierarquia ponderada das vias, a idade média da população e a densidade do somatório da população com o número de empregos. A variável dependente utilizada foi a média da densidade dos acidentes nos anos de 2008 a 2010. Os modelos testados foram os modelos de regressão múltipla, os modelos lineares generalizados com distribuição de Poisson e binomial negativa e os modelos espaciais *Spatial Autoregressive* (SAR) e *Conditional Autoregressive* (CAR). Todos estes modelos ainda foram testados em diferentes regimes espaciais gerados no programa de regionalização espacial REDCAP. O modelo que apresentou os melhores resultados foi o modelo de regressão múltipla sem regimes espaciais.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

SPATIAL ANALYSIS OF TRAFFIC ACCIDENTS WITH REGIONAL VARIABLES: METHODOLOGY AND APLICATION IN THE CITY OF RIO DE JANEIRO

Marcos de Meneses Rocha

March/2015

Advisor: Carlos David Nassi

Department: Transportation Engineering

This thesis presents a methodology for the spatial analysis of georeferenced traffic accidents in the city of Rio de Janeiro, in the context of areal variables and that takes in to account the specifities of spatial data, such as spatial dependence, spatial heterogeneity, the Modifiable Areal Unit Problem (MAUP), ecological fallacy and edge effects. The methodology consists of three main stages: acquisition, understanding and modeling the spatial distribution of accident data. The most explanatory variables were weighted hierarchy of roads, the average age of population and the density of the sum of population and the number of jobs. The dependent variable used was the average density of accidents in the years 2008 to 2010. The non-spatial models used were the multiple regression models and generalized linear model with Poisson distribution and negative binomial models. The spatial models were Spatial Autoregressive (SAR) and Conditional Autoregressive (CAR). All these models were also tested at different spatial regimes generated in the space regionalization program REDCAP. The best model was the multiple regression model without spatial regimes.

## ÍNDICE DO TEXTO

## LISTA DE FIGURAS

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS E SIGLAS

ALK - *Average LinKage*
BRAT - Boletim de Registro de Acidentes de Trânsito
CAR – *Conditional Autoregressive*
CLK - *Complete LinKage*
CET-Rio - Companhia de Engenharia de Tráfego da Cidade do Rio de Janeiro
CURE – *Cumulative Residuals*
DNER – Departamento Nacional de Estradas de Rodagem
DNIT - Departamento Nacional de Infraestrutura de Transportes
EJ - *Environmental Justices Areas*
FETRANSPOR - Federação das Empresas de Transportes de Passageiros do Estado do Rio de Janeiro
GEIA - Grupo Executivo da Indústria Automobilística
GPS – *Global Positioning System*
IMD - *Index of Multiple Deprivation*
LISA – *Local Indicators of Spatial Association*
LSOA - *Lower Super Output Area*
MAD - *Mean Absolute Deviation*
MAUP - *Modifiable Areal Unit Problem*
ML – Multiplicadores de Lagrange
MLG- Modelos Lineares Generalizados
NRMS - *Normalized Root Mean Square Deviation*
MSPE - *Mean Squared Predictive Error*
PIB – Produto Interno Bruto
RAIS/MTE - Relatório Anual do Trabalho e do Emprego fornecido pelo Ministério do Trabalho e Emprego
REDCAP - *Regionalization with dynamically constrained agglomerative clustering and partitioning*
RMS - *Root Mean Square Error ou Root Mean Square Deviation*
SAR – *Spatial Autoregressive*
SIAT-FOR - Sistema de Informações de Acidentes de Trânsito de Fortaleza
SLK - S*ingle LinKage*
STPP - *Surface Transportation Policy Project*
UTM -*Universal Transversa de Mercator*
ZUS - *Zones Urbaines Sensibles*

# 1. INTRODUÇÃO

Desde 16 de junho de 1956, dia considerado como o marco do nascimento da indústria automobilística no Brasil, quando Juscelino Kubitschek, cinco meses após sua posse, assinou o Decreto nº. 39.412, criando o Grupo Executivo da Indústria Automobilística (GEIA), a indústria automobilística vem se desenvolvendo gradativamente no Brasil. Em 2011 atingiu-se a marca de 26 montadoras de veículos e 53 fábricas instaladas em 9 estados e 39 municípios, com uma produção de cerca de 3,4 milhões de veículos, segundo dados do Anuário Estatístico da Indústria Automobilística Brasileira de 2012 (ANFAVEA, 2012), colocando o país entre os maiores produtores mundiais de veículos.

A grande oferta de automóveis no Brasil, aliada às melhoras de condição de renda e de oferta de crédito vêm fazendo com que a frota de automóveis sofra um salto nos últimos dez anos. Segundo dados do DENATRAN(2012), a frota de automóveis em algumas cidades mais que dobrou no período, como é o exemplo de Manaus, que teve um aumento da frota da ordem de 160% no período de 2001 a 2011. No caso do Rio de Janeiro, o aumento no mesmo período foi de 57%. No mesmo período, a frota de motocicletas aumentou cerca de 300%, segundo o Anuário Estatístico da Associação Brasileira dos Fabricantes de Motocicletas, Ciclomotores, Motonetas, Bicicletas e Similares (ABRACICLO, 2012).

Por outro lado, desde a década de 50, a população brasileira cresceu quase 4 vezes, atingindo valores em torno dos 190 milhões de habitantes, segundo o CENSO demográfico de 2010 produzido pelo IBGE, sendo que cerca de 84% da mesma vive nas cidades.

Como seria de se esperar a partir do aumento da frota e da população urbana, vem ocorrendo um crescente aumento nos acidentes de trânsito no Brasil. Segundo MORAIS NETO *et al.* (2012), o risco de morte por acidentes de transporte terrestre aumentou de 18 para 22,5 para cada 100 mil habitantes no Brasil, no período de 2000 a 2010.

A presente tese busca contribuir para o melhor entendimento sobre os acidentes ocorridos em vias urbanas a partir de uma metodologia de análise espacial dos dados de acidentes, utilizando como estudo de caso a cidade do Rio de Janeiro.

A primeira parte da tese é constituída por esta breve introdução ao assunto e de mais três seções: objetivo e originalidade da tese, relevância e justificativa e, por fim, estrutura do trabalho escrito.

## 1.1. Objetivo e originalidade

O objetivo desta tese é propor uma metodologia de análise dos acidentes de trânsito georreferenciados que considere as especificidades dos dados geográficos, empregando variáveis agregadas em áreas e relacionadas à geometria e conectividade das vias, às características demográficas, socioeconômicas e de uso do solo, verificando sua aplicabilidade no município do Rio de Janeiro

Conforme será visto em detalhes na revisão bibliográfica, o uso de dados georreferenciados é crescente na modelagem de acidentes. No entanto, trabalhar dentro deste ambiente envolve conhecer diversas premissas, as quais não costumam ser consideradas, tais como dependência espacial, heterogeneidade espacial, a questão do *Modifiable Areal Unit Problem* (MAUP), falácia ecológica e efeito das bordas. Quando presentes na bibliografia de acidentes, são feitas considerando alguns dos aspectos isoladamente e não de forma conjunta e integrada.

Por outro lado, o fato de se espacializar o fenômeno em estudo abre espaço para o emprego de diversas técnicas de visualização dos dados e de análise espacial, as quais têm potencial de trazer uma visão mais rica sobre o comportamento dos acidentes que não somente a oriunda da interpretação de modelos e testes estatísticos.

A originalidade da presente tese estaria em apresentar uma forma estruturada de fazer a análise espacial de acidentes de trânsito que integre a análise do fenômeno geográfico, comumente empregada na Geografia e visualizada por meio de mapas, com a análise estatística de dados espaciais. A primeira concepção está concentrada nas etapas de aquisição e compreensão da distribuição espacial e a segunda na de modelagem, embora a primeira esteja presente em diversas fases da segunda.

Quando da aplicação da metodologia, algumas hipóteses serão testadas na região de estudo. São elas:

1) A hierarquia das vias pode ser empregada como indicadora de exposição ao risco de acidentes, tendo em vista não serem disponibilizadas as informações sobre fluxo de veículos no Rio de Janeiro;

2) Os modelos espaciais apresentam resultados superiores aos modelos não espaciais equivalentes, devido à consideração da dependência espacial;

3) Os modelos estatísticos que consideram a divisão da região de estudo em regimes espaciais apresentam melhores resultados que os equivalentes cujas variáveis explicativas

possuem coeficientes constantes; e

4) A divisão da região em quantidades diferentes de áreas de agregação visando contemplar o MAUP altera consideravelmente os resultados dos modelos.

### 1.2. Relevância e justificativa

Segundo o relatório da Organização das Nações Unidas sobre estado global da segurança viária (WHO, 2009), as mortes por acidentes de trânsito são a maior causa de morte na população entre 15 e 29 anos e serão a 5ª maior causa de morte no mundo considerando todas as idades em 2030, sendo que o Brasil consta como o 5º país com maior quantidade de acidentes de trânsito no mundo.

Por outro lado, os custos envolvidos com os acidentes também revelam valores expressivos. Segundo o relatório sobre impactos sociais e econômicos dos acidentes de trânsito nas rodovias brasileiras do Instituto de Pesquisa Econômica Aplicada – (IPEA, 2006), o custo dos acidentes nas rodovias brasileiras atingiu cerca de 1,2% do Produto Interno Bruto (PIB) brasileiro em 2005. No relatório do mesmo Instituto, elaborado em 2003, sobre impactos sociais e econômicos dos acidentes de trânsito nas aglomerações urbanas (IPEA, 2003), o custo dos acidentes nas áreas urbanas atingiu um valor que correspondia a aproximadamente 0,5% do PIB brasileiro em 2001.

Ainda nesse sentido e a partir de informações de inquérito obtidas pelo Ministério da Saúde sobre violências e acidentes no Brasil, verificou-se que os acidentes provocados pelo trânsito foram responsáveis pela segunda maior quantidade de atendimentos por acidentes em unidades de saúde do Brasil, perdendo somente para as quedas, impactando sobremaneira nos custos de saúde.

Tendo em vista tal realidade, a Organização das Nações Unidas elegeu o período de 2011-2020 como sendo a Década de Ação pelo Trânsito Seguro, na qual governos de todo o mundo se comprometem a tomar novas medidas para prevenir os acidentes no trânsito. No Brasil, o governo brasileiro vem realizando diversas ações, dentre as quais lançou o Projeto Vida no Trânsito, onde se busca, dentre outros objetivos, o de identificar fatores de riscos e grupos de vítimas mais vulneráveis a acidentes de transportes terrestres.

No caso do estado do Rio de Janeiro, uma análise dos dados de acidentes faz-se necessária tendo em visto a mudança a partir de 2012 da forma de registro dos acidentes, passando do registro no local do acidente para uma forma eletrônica, o que deve gerar descontinuidade nas estatísticas de acidentes. Essa forma eletrônica de registro de

acidentes denomina-se e-BRAT e deve ser feita pelos envolvidos nos acidentes em caso de acidentes com somente danos materiais, a partir da submissão do pedido de confecção do registro à Policia Militar diretamente na Internet e não mais na presença do policial no local do acidente.

Nesse contexto, a presente tese tem como **primeira justificativa** a necessidade de se criar uma metodologia para a análise espacial de acidentes de trânsito que considere as especificidades dos dados geográficos de acidentes;

Como **segunda justificativa**, tem-se a de identificar alguma variável indicadora da exposição ao risco de acidentes para os locais onde não existem de dados de fluxo de veículos;

Por meio da revisão bibliográfica, tem-se como **terceira justificativa** a de se poder verificar se algumas variáveis explicativas adotadas na bibliografia se aplicam à cidade do Rio de Janeiro;

Como **quarta justificativa**, tem-se a de verificar o comportamento da dependência e heterogeneidade espaciais neste local e propor modelos que possam auxiliar na sua correção.

Espera-se que esta metodologia possa fornecer subsídios para uma melhor compreensão da distribuição espacial dos acidentes de trânsito e para a modelagem dos acidentes enquanto dados geográficos na cidade do Rio de Janeiro.

Tendo em vista as peculiaridades de cada região, não se tem expectativa de que os parâmetros do modelo possam ser aplicados diretamente em outros locais sem que tenham sido feitas as análises que antecedem a etapa de modelagem propriamente dita. Até mesmo para o mesmo local em um período diferente, espera-se que seja feita uma análise crítica do comportamento das variáveis nesse novo instante de tempo.

### 1.3. Estrutura do trabalho

O presente trabalho está dividido em seis capítulos.

O primeiro reúne as informações que introduzem o tema da tese, apresentam o objetivo, originalidade, relevância e justificativa do trabalho, bem como a estrutura do texto.

No segundo capítulo, é feita uma análise dos principais conceitos relativos à segurança viária, assim como a revisão bibliográfica das principais variáveis explicativas que vem sido associadas às análises estatísticas de acidentes nos últimos anos, para o caso

dos modelos de frequência de acidentes. Em seguida, apresenta-se uma revisão bibliográfica sobre a análise estatística de acidentes no Brasil, restrita principalmente à análise exploratória e modelagem dos acidentes. Por fim, apresenta-se uma revisão sobre a análise de dados de acidentes agregados em área, a qual será empregada na tese.

No terceiro capítulo, é feita uma exposição sucinta sobre os diversos conceitos que envolvem a análise espacial, indispensáveis para a compreensão da metodologia. São eles as características dos dados geográficos, SIG e análise espacial, análise visual, análise exploratória, modelagem estatística, verificação da qualidade do ajuste e validação do modelo.

O quarto capítulo está reservado para a metodologia proposta, a qual está dividida em três grandes etapas: aquisição, compreensão da distribuição espacial e modelagem dos dados de acidentes. A etapa de aquisição compreende a de caracterização da área de estudo, coleta e preparação dos dados. Na etapa de compreensão estão a verificação dos efeitos de bordas, a análise visual e a análise exploratória dos acidentes. Na etapa de modelagem estão a seleção das variáveis explicativas, a calibração dos dados estatísticos, a análise da heterogeneidade espacial, a geração e análise dos dados para verificar a questão do MAUP e a validação dos modelos.

No quinto capítulo, são apresentados os resultados e realizada ao mesmo tempo a síntese dos mesmos para o caso da região de estudo.

Finalmente, o sexto capítulo apresenta as conclusões retiradas a partir dos resultados obtidos.

## 2. SEGURANÇA VIÁRIA

O presente capítulo traz uma revisão bibliográfica dos principais conceitos de segurança viária e das variáveis explicativas e modelos empregados na previsão de acidentes.

### 2.1 Conceitos de segurança de tráfego

Segundo o IPEA (2006), o qual utiliza as definições adotadas na Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde – Décima Revisão (CID-10) da Organização Mundial da Saúde (OMS), acidente é "um evento independente do desejo do homem, causado por uma força externa, alheia, que atua subitamente (de forma inesperada) e deixa ferimentos no corpo e na mente".

Segundo o IPEA (2006), acidente de trânsito é "todo acidente com veículo ocorrido em via pública". A via pública seria "a largura total entre dois limites de propriedade de todo o terreno ou caminho aberto ao público, quer por direito, quer por costume, para a circulação de pessoas ou bens de um lugar para outro". Dentro das vias públicas, tem-se a pista ou leito de via como sendo "a parte da via pública que é preparada, conservada e, habitualmente, usada para o trânsito de veículos".

Considerando-se que o Código de Trânsito Brasileiro define trânsito como sendo "a movimentação e imobilização de veículos, pessoas e animais nas vias terrestres", os acidentes de trânsito devem neste contexto envolver inclusive os acidentes ocorridos com os pedestres nas calçadas.

Segundo o IPEA (2006), pedestre "é toda pessoa envolvida em um acidente mas que, no momento em que este ocorreu, não estava viajando no interior ou sobre um veículo motor, trem em via férrea, bonde, veículo de tração animal ou outro veículo, ou sobre bicicleta ou animal".

Segundo a ABNT (1993), condutor é "toda pessoa que conduza um veículo automotor, ou de outro tipo, incluindo os ciclos, ou que gire por uma via, cabeças de gado isoladas, rebanho, bando ou manadas, ou animais de tiro, carga ou sela".

Segundo a ABNT(1993), vítima de acidente de trânsito é "toda pessoa que sofre lesões físicas e/ou perturbações mentais, em razão de acidente de trânsito, independente de culpa civil ou penal".

Os acidentes de trânsito, por sua vez, podem ocorrer nos trechos de vias públicas, no

interior dos quarteirões ou nas áreas em que duas ou mais vias se cruzam ou se unificam denominadas de interseção.

## 2.2 Causas de acidentes

Os acidentes podem ocorrer devido a um fator ou pela combinação de diversos fatores. A bibliografia costuma mencionar como principais fatores que influenciam a ocorrência dos acidentes os fatores humanos, viário-ambientais e veiculares. Segundo GOLD (1998), os fatores contribuintes para os acidentes são: fatores humanos, fatores relativos ao veículo, fatores relativos à via ou meio ambiente e ambiente construído e fatores institucionais ou sociais.

Segundo ALMEIDA (2011), o fator humano diz respeito ao comportamento, à educação, incluindo nessa a educação no trânsito, a habilidade na condução do veículo, bem como as condições físicas e psicológicas do indivíduo. No caso das condições físicas podem ser incluídas, dentre outras, o cansaço e o consumo de álcool ou drogas. Quanto às condições psicológicas, tem-se a distração ou riscos causados por tensões ou emoções. As características demográficas, tais como sexo e faixa etária, possuem grande influência nesse sentido.

Quanto ao fator veicular, tem-se as especificações do veículo que podem influenciar nos acidentes, tais como potência, capacidade de frenagem, uso de equipamentos de segurança como *airbags*, por exemplo, bem como as condições do mesmo tal como estado dos pneus, funcionamento das luzes de freio, etc.

Nos fatores viário-ambientais estão as características da via, tais como tipo de revestimento, largura e aqueles referentes às características de projeto, como geometria, superelevação, etc., bem como a sinalização ao longo das mesmas. Nas características ambientais, estão enquadradas as condições climáticas, como a existência de chuva, neblina, luminosidade, etc.

Os fatores institucional/social seriam aqueles referentes à regulamentação e policiamento, incluindo o estabelecimento das condições de circulação nas vias e a sua fiscalização.

Além desses fatores, mais focados nas condições específicas dos acidentes, pode-se elencar outros fatores associados ao local onde ocorreu acidente, que não sejam referentes às condições da via. Como exemplo, tem-se as características socioeconômicas e demográficas dos moradores do bairro onde ocorrera o acidente, bem como o uso do solo.

No que diz respeito ao uso do solo, pode-se levar em consideração fatores como o relevo, o fato da região possuir um uso do solo predominantemente residencial, comercial, industrial ou uso misto, a de apresentar grandes áreas de lazer públicas ou áreas verdes, a existência de estabelecimentos com venda de bebidas alcoólicas, entre outros fatores.

### 2.3 Exposição ao risco

O conceito de exposição ao risco ajuda a entender o risco que cada indivíduo está sujeito a partir do momento em que decide transitar de um local para outro, seja a pé ou por meio de um veículo para o deslocamento. Segundo RAFORD e RAGLAND (2003, *apud* THAKURIAH e COTTRILL, 2008), exposição pode ser definida como a taxa de contato com um potencial agente ou evento perigoso sendo a exposição de um pedestre ou veículo a taxa de contato destes com uma potencial situação envolvendo veículos. LASSARE *et al.* (2012a) empregam o conceito de exposição ao risco utilizado na epidemiologia ambiental, adotando o conceito da Academia Nacional de Ciência (1991) a qual define exposição como "um evento que ocorre quando existe um contato do homem com a fronteira de um meio ambiente onde está presente um contaminante de uma específica concentração por um intervalo de tempo". Nesse sentido, existe um contato virtual entre um usuário de uma via, seja ele um veículo ou pedestre, e uma certa atmosfera gerada pelo tráfego, sendo que a qualidade dessa atmosfera depende da presença de contaminantes que correspondem ao movimento de veículos descrito pelo volume de tráfego e a velocidade dos veículos.

Segundo LASSARE *et al.* (2012b), o tempo gasto no trânsito costuma ser conhecido pelos especialistas em segurança viária como indicador de exposição ao risco. Uma grandeza comumente empregada nesse sentido é a distância diária percorrida pelos veículos. Quanto aos pedestres, o tempo de exposição pode ser definido como o tempo gasto pelo mesmo para atravessar uma determinada via de certa largura com certa velocidade de caminhada, considerando neste caso que o pedestre estaria em risco somente no momento em que estaria em contato com uma via com fluxo de veículos. A tendência é que em locais onde as pessoas estejam mais expostas ao risco, ocorram mais acidentes.

## 2.4 Coleta dos dados de acidentes

No registro dos acidentes de trânsito, as informações de acidentes costumam ser obtidas por policial, o qual preenche os chamados Boletins de Ocorrência. No Rio de Janeiro, tais registros são chamados de Boletins de Registro de Acidentes de Trânsito (BRAT), para o caso de ocorrerem danos materiais e/ou danos físicos. Quando ocorrem danos físicos, independente da gravidade, costuma-se preencher os Registros de Ocorrência nas delegacias da Polícia Civil. Quanto aos danos materiais, nos anos cujos dados foram utilizados na pesquisa ainda se preenchiam os BRAT no local do acidente. Conforme mencionado anteriormente, a partir de 2012, no estado do Rio de Janeiro, os envolvidos em acidentes de trânsito passaram a preencher o registro diretamente nas Unidades da Polícia Militar e posteriormente por meio eletrônico (eBRAT).

Conforme SOUZA (2011), as condições em que é preenchido o boletim, bem como o tipo de profissional que o preenche pode alterar a qualidade do preenchimento. Por exemplo, quando se tem um acidente somente com danos materiais, o registro no local do acidente pode conter imprecisões devido ao estado emocional dos envolvidos. No caso de quando ocorrem danos físicos aos envolvidos, o mesmo pode ser preenchido por um profissional da área de saúde no local onde os envolvidos foram removidos, o qual costuma estar mais focado no estado do paciente do que nas condições do acidente.

A qualidade com que se preenchem tais informações é de fundamental importância para a adequada compreensão do acidente. No contexto de uma modelagem estatística, podem ser empregadas diretamente como variáveis explicativas ou podem ser indicadoras de variáveis que poderão ser obtidas a partir de outras fontes como, por exemplo, censos ou estatísticas. No Brasil, infelizmente, o registro de acidentes de trânsito costuma ser feito muitas vezes de forma inadequada, o que acaba por gerar dificuldade para se aproveitar tais informações. Somado a isso, o fato de se realizá-los em papel faz com que muitas vezes ao serem inseridos em banco de dados, corra-se o risco de se incorrer em erros de leitura, interpretação ou de digitação. O preenchimento dos dados em um formulário digital poderia reduzir tais erros grosseiros. QUEIROZ (2004b) aborda diversos exemplos destas dificuldades encontradas quando da introdução dos dados no Sistema de Informações de Acidentes de Trânsito de Fortaleza(SIAT-FOR).

Quanto à posição onde ocorreram os acidentes, costuma ser obtida pelo endereço e/ou ponto de referência. No caso do Brasil, é comum que os boletins contenham endereços incompletos, faltando a indicação do número do endereço ou utilizando pontos

de referência que já não estejam presentes no local anos depois, como o nome de um estabelecimento comercial. A melhor forma de atenuar os problemas de localização dos acidentes é associando os mesmos às coordenadas referenciadas a um sistema geodésico. Nesse sentido, a forma mais empregada de coletar tais coordenadas seria utilizando rastreadores de satélites GPS (*Global Positioning System*), o qual pode fornecer coordenadas com alguns erros de posição, dependendo da época e do local. Na época em que foram coletados os dados de acidentes utilizados na pesquisa, o sistema GPS apresentava erros de posicionamento em torno de 8 metros, com nível de confiança de 95%, segundo o Serviço de Posicionamento Padrão do GPS, Departamento de Defesa Americano. Tais erros poderiam ser ainda maiores caso fossem inseridos outros erros, tal como o de multicaminhamento, muito comum em áreas urbanas, decorrente da mudança da trajetória do sinal que vem do satélite devido à reflexão em prédios, massas d´água e outros obstáculos. Mais detalhes sobre posicionamento com GPS podem ser encontrados em MONICO (2000).

## 2.5 Gerenciamento da segurança de trânsito

Segundo CARDOSO (2006), o gerenciamento da segurança viária envolve as análises e todas as ações que possam ser tomadas com vistas a aumentar a segurança viária. Para NODARI (2003), o gerenciamento da segurança viária pode ocorrer de forma reativa ou corretiva, bem como de forma pró-ativa ou preventiva. No gerenciamento de forma reativa, procura-se solucionar os problemas de acidentes a partir da análise dos locais com maior incidência dos mesmos. Como exemplo, tem-se a identificação de regiões com grande concentração de acidentes. No caso do gerenciamento preventivo, busca-se identificar situações e/ou locais que ofereçam risco potencial de acidentes, atuando-se antes que o acidente venha a ocorrer. Como exemplo, têm-se as técnicas de auditoria de segurança viária e análise de conflito de tráfego.

O DENATRAN (1987) ainda menciona duas maneiras de se tentar solucionar os problemas de acidentes de trânsito: no nível macro, com programas mais abrangentes de educação, fiscalização, etc. e no nível micro, por meio de estudos dos locais com maior incidência de acidentes.

A análise espacial dos acidentes de trânsito, empregada no contexto desta tese, utiliza dados estatísticos de acidentes podendo ser considerada como sendo como um exemplo de ação reativa. A partir do momento em que se aplica a metodologia em outras regiões

ou em um segundo momento no mesmo local, passa a ser uma ferramenta com potencial para o gerenciamento preventivo dos acidentes.

## 2.6 Identificação dos locais dos acidentes

Visando aumentar o conhecimento sobre os locais com maior incidência de acidentes, procurou-se associar a descrição dos mesmos, em um primeiro momento, à rede viária, cujos nós seriam as interseções das vias (neste caso sem considerar os nós referentes ao final das ruas) e os arcos representariam os trechos de vias.

No Brasil, as interseções e trechos de vias foram chamados de locais críticos. Segundo DENATRAN (1987), os pontos críticos são os locais com maiores taxas de acidentes de trânsito devendo, portanto, receber prioridade no seu tratamento com a finalidade de eliminar ou, ao menos, reduzir tal taxa. A identificação de pontos críticos pelo DENATRAN (1987) é feita por meio do cálculo de taxas de acidentes, em unidade-padrão de severidade, cujos valores são diferentes para as interseções e para os trechos de via, em função do volume médio diário de veículos passando por cada aproximação (no caso das interseções) ou por km de trecho de via. As unidades-padrão de severidade são pesos atribuídos aos acidentes, no caso 1 para acidentes somente com danos materiais, 5 para acidentes com feridos e 13 para acidentes com mortos. Tais pesos estão baseados nos custos envolvidos em cada um dos tipos de acidentes, levantados pelo então Departamento Nacional de Estradas de Rodagem (DNER), atual Departamento Nacional de Infraestrutura de Transportes (DNIT). Tais pesos seriam mais adequados para o caso de estradas e com custos de 1980. No caso das vias urbanas, segundo IPEA (2003, *apud* CARDOSO, 2006) os pesos mais adequados seriam de 1 para acidentes somente com danos materiais, 5 para acidentes com feridos e 44 para acidentes com mortos.

Na literatura, os pontos críticos vêm sendo conhecidos com os termos *hotspots* (ANDERSON, 2009) ou como *blackspots* (FLAHAUT *et al.*, 2003). ANDERSON (2009) inclusive cria uma classificação para os locais com maior índice de acidentes em função da escala em que se esteja trabalhando, indo do *blackspots*, passando pelos aglomerados de *blackspots* chamados pelo autor de *clusters* de acidentes até chegar nos grupos de acidentes *(*do inglês, *groups).* (Figura 1).

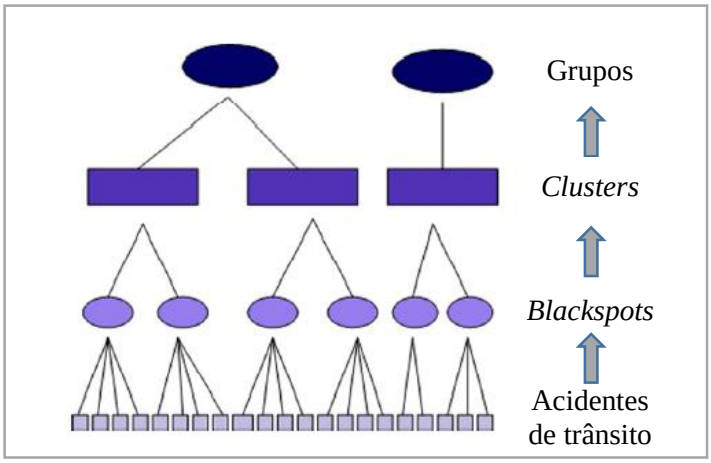


Figura 1Classificação dos locais dos acidentes de acordo com a escala
Fonte: ANDERSON (2009)

Um modo comum de construir pontos críticos tem sido a de atribuir a células de uma grade regular valores numéricos associados ao número de acidentes ocorridos dentro da mesma ou na sua vizinhança. Os locais de risco para acidentes (chamados de *hotspots ou blackspots*) são obtidos a partir da agregação destas células, podendo ter tamanhos variados. Regiões que apresentam concentração de *blackspots* adjacentes são comumente denominadas de *black zones*. QUEIROZ *et al.* (2004a) identificam agrupamentos de acidentes de trânsito na cidade de Fortaleza pelo critério dos vizinhos mais próximos e apresenta-os por meio da elipse de desvio padrão. ANDERSON (2009) aplica a técnica de *K-means* sobre dados de acidentes de trânsito da cidade de Londres, sendo esta uma técnica em que os agrupamentos são determinados a partir da minimização dos valores das variâncias entre os dados dos *blackspots* e os centroides dos *clusters*.

Outra forma de estudar os acidentes é associando os mesmos às áreas de agregação, como é o caso de zonas de tráfego, bairros, dentre outras, como é o caso de MIRANDA-MORENO *et al.* (2011), os quais empregaram como áreas de agregação *buffer zones* de 50, 150, 400 e 600m em torno de interseções semaforizadas. A possibilidade de correlacionar os acidentes com outros dados, como é o caso daqueles agregados em setores censitários, vem tornando essa forma de analisar a distribuição espacial dos acidentes cada vez mais utilizada.

### 2.7 Variáveis explicativas associadas aos acidentes

Tendo em vista que a presente metodologia converge para a construção de modelos estatísticos, nos itens seguintes será feita uma revisão bibliográfica das principais

variáveis explicativas, bem como dos modelos estatísticos propriamente ditos, empregados nos últimos anos na análise de dados de acidentes de trânsito.

As variáveis explicativas aqui apresentadas serão aquelas obtidas de outras fontes que não dos boletins policiais, os quais contemplam as características dos ocupantes dos veículos, como idade e sexo, as características do veículo, as características de construção das vias, o horário e o dia da semana, bem como as condições meteorológicas no momento do acidente. Serão apresentadas, no lugar, variáveis provenientes de outras fontes de informação e que a bibliografia vem apontando como sendo correlacionadas com os acidentes de trânsito em diversos locais do mundo.

Na prospecção dos fatores que possam estar associados aos acidentes de trânsito é importante conhecer o que torna uma região mais propensa a tais tipos de acidentes. A exposição ao risco de uma dada região é aqui entendida como um somatório de riscos de acidentes individuais de quem circula nesta área, podendo ser associada à quantidade de pedestres que circulam na região, ao volume de tráfego e à velocidade máxima nas vias (MIRANDA-MORENO *et al.*, 2011). Neste item será feito um breve comentário sobre cada um desses três itens, bem como sobre as variáveis explicativas associadas à geometria e conectividade das vias, características demográficas, socioeconômicas e de uso do solo.

### 2.7.1 Fluxo de veículos

O fluxo de veículos pode obtido a partir da contagem volumétrica dos veículos nos trechos das vias ou interseção das mesmas, que pode ser feita diretamente de forma manual ou automatizada por meio de contadores, bem como a partir dos Instrumentos de Medição de Velocidade Autônomos, conhecidos como radares, pardais ou lombadas eletrônicas. No caso do Brasil, tais contagens costumam ser feitas nas vias urbanas para estudos específicos visando a construção de alguma benfeitoria no local, como passarelas, viadutos, alargamento de calçadas, etc., ou para o planejamento do trânsito para algum evento.

O fluxo de veículos, sendo um indicador de exposição ao risco, tem um grande impacto no número de acidentes. Nesse sentido, MIRANDA-MORENO *et al.* (2011) mostraram que uma redução de 30% no volume de veículos nas interseções de sua área de estudo, diminuíam em 50% o risco médio dos pedestres e em 35% dos acidentes com danos aos pedestres. BRAGA *et al.* (2005) analisaram a relação entre a exposição ao

tráfego e os acidentes na cidade do Rio de Janeiro.

Uma variável que pode funcionar como um indicador das vias com maior ou menor fluxo de veículos e que vem sendo utilizada por diversos autores é a hierarquia das vias. MIRANDA-MORENO *et al.* (2011), por exemplo, empregaram a classificação de vias do tipo expressas, arteriais e locais, encontrando nas arteriais o maior risco de acidentes envolvendo os pedestres. No caso da cidade do Rio de Janeiro, a Companhia de Engenharia de Tráfego (CET-Rio) adota uma hierarquização viária baseada no Plano Diretor Decenal da cidade do Rio de Janeiro, o qual estabelece 5 (cinco) centros de alcance metropolitano e 14 (quatorze) centros de alcance municipal. Os centros metropolitanos são Centro, Copacabana, Barra da Tijuca, Madureira e Campo Grande. Os centros de alcance municipal são Estácio, Tijuca, Ipanema, Botafogo, Penha Circular, Ramos, Méier, Bonsucesso, Irajá, Ilha do Governador, Taquara/Tanque, Pavuna, Campo Grande e Bangu. A Tabela 1 apresenta a hierarquia das vias, bem como a principal função das mesmas.

Tabela 1 Hierarquia das vias adotada pela CET-Rio

| **Classificação das vias** | **Função** |
|---|---|
| Estruturais | Ligações rápidas para atender deslocamentos de longa distância e alto volume de veículos |
| Arteriais primárias | Ligação dos centros de alcance metropolitanos e destes com as estruturais |
| Arteriais secundárias | Ligação dos centros de alcance municipal e destes com os centros de alcance metropolitanos destes com as estruturais e arteriais |
| Coletoras | Coleta e distribuição de tráfego interno dos bairros e alimentação das arteriais |
| Locais | Acesso direto às residências, comércio e indústrias, com tráfego exclusivamente local |

Os fluxos de pedestres e de veículos em um dado local serão então aqui considerados de maneira indireta, ou seja, a partir da quantificação de fatores que possam influenciar no volume de veículos e de pedestres de um dado local. Além disso, serão aqui divididos os fluxos entre aqueles decorrentes de motivações que possam exceder os limites geográficos da região de estudo, aqui chamados de globais, e aqueles decorrentes de motivações situadas dentro ou nas proximidades dos limites geográficos da região de estudo, denominados de locais.

Como exemplo das motivações globais, tem-se o fato da região estar situada entre

regiões da cidade com grande população e oferta de serviços. No caso dos veículos, será aqui considerada a hierarquia das vias como sendo um indicador nesse sentido. No caso dos pedestres, a acessibilidade aos transportes públicos pode ser um indicador com motivações globais, tendo em vista que a região pode ser atravessada por transporte público cuja origem e destino estejam fora da região de trabalho ou podem ser empregados como integrações para outro que se destinem a outras regiões.

No caso das motivações locais, tem-se a população local e toda a oferta de empregos, de serviços e de opções de lazer disponíveis na região, que fazem com que exista um maior ou menor fluxo de pessoas circulando na mesma.

### 2.7.2 Fluxo de pedestres

Nos últimos anos, grande parte dos estudos de modelagem de acidentes de tráfego em ambiente de SIG vêm sendo realizados empregando a colisão de veículos com pedestres (UKKUSURI *et al.*, 2012, MIRANDA-MORENO *et al.*, 2011, HA e THILL, 2011, PULUGURTHA e SAMBHARA, 2011, COTTRILL e THAKURIAH, 2010, BLAZQUEZ e CELIS, 2013, AZIZ *et al.*, 2013, NOLAND *et al.*, 2013, WANG e KOCHELMAN, 2013, ELIAS e SHIFTAN, 2014).

A contagem do fluxo de pedestres, da mesma forma que a de veículos, é feita de forma escassa no Brasil e comumente para fins específicos. Como forma de se obter indicadores de fluxo de pedestres, tem-se recorrido a diversas variáveis demográficas e econômicas, tais como população e número de empregos, a acessibilidade do local, o uso do solo e o próprio design das vias.

Como forma de melhor compreender o comportamento dos pedestres, o qual pode refletir no fluxo dos mesmos, será feito um breve comentário sobre os principais fatores que podem incentivar ou não esse modo de transporte.

A capacidade de se percorrer a pé um determinado local é um importante fator que faz com que as pessoas possam caminhar com mais frequência e por uma maior extensão em um determinado local. Como um primeiro fator a ser considerado, tem-se a qualidade das calçadas disponibilizadas ao pedestre. Nesse sentido, LANDIS *et al.* (2001) descrevem três medidas para se medir a performance do ambiente das calçadas. São eles: 1) a capacidade física da calçada, 2) a qualidade do ambiente para caminhar e 3) a percepção do pedestre de segurança e conforto em relação ao tráfego de veículos motorizados. Conforme PETRITSCH *et al.* (2007), a medida da capacidade física da

calçada é adequada para avaliar calçadas existentes ou projetadas, não sendo adequada como forma de valorar e priorizar calçadas que precisem ser remodeladas. Quanto à qualidade do ambiente, está associado ao prazer que o pedestre tenha de percorrer a calçada e visa criar condições mais convidativas para se caminhar. No que diz respeito à percepção de segurança/conforto em relação ao tráfego de veículo, LANDIS *et al.* (2001) mencionam os seguintes fatores que influem no senso de segurança/conforto dos pedestres: presença de calçada, separação lateral do tráfego de veículos, existência de barreiras ou algum tipo de amortecimento entre os pedestres e o fluxo de veículos, tipo e volume de veículos, velocidade dos veículos e frequência e volume de entradas de automóveis. ARAUJO e BRAGA(2008) fizeram uma avaliação qualitativa de travessias de pedestres a partir da determinação do nível de serviço medido pela entrevista com os usuários das mesmas e comparado com o nível de serviço quantitativo calculado de acordo com o HCM (*Highway Capacity Manual*) de 2000.

O FHWA (1999) menciona como áreas críticas para pedestres também a proximidade do destino, a continuidade das ruas, o tamanho médio dos quarteirões e a declividade.

LASSARRE *et al.* (2012b) mencionam diferentes indicadores sobre o comportamento dos pedestres, incluindo os macroscópicos e os microscópicos. Os macroscópicos envolvem indicadores tais como número de viagens, número de faixas de pedestres, tempo de caminhada e distância viajada. No entanto, afirma que os mesmos não são tão precisos a ponto de avaliar o comportamento dos pedestres e a exposição ao risco. KELLY *et al.* (2011), mencionam diversos métodos microscópicos que vêm sendo utilizados para avaliar o ambiente dos pedestres, destacando-se aqueles que procuram verificar a capacidade de se caminhar em uma determinada rota, bem como aqueles que determinam as preferências do mesmo ao percorrer tal rota. A preferência que vem mais sido explorada nos últimos anos é a escolha do momento de atravessar as vias. PAPADIMITRIOU (2012) apresenta uma proposta de modelagem no padrão de travessia de pedestres em vias urbanas, observando que existe uma tendência dos pedestres atravessarem as vias no início das viagens mais curtas e adiarem as travessias quando se percorre maiores distâncias, principalmente para os pedestres que andam com maior velocidade. O mesmo autor ressalta ainda que os pedestres tendem a atravessar nos semáforos nas vias com maior fluxo. No entanto, naquelas com menor fluxo e somente um sentido, o pedestre tende a atravessar no meio do quarteirão.

### 2.7.3 Velocidade das vias

A velocidade em que os veículos transitam possui impacto no número de acidentes, tendo em vista em que aumenta a possibilidade de se ter colisões entre os veículos e entre os mesmos e os pedestres, na medida em que não se tem tempo hábil para se desvencilhar de situações de risco. Tem impacto também na severidade dos acidentes, embora a fatalidade tende a aumentar em regiões mais rurais (JOHNSON e LU, 2011, DONALDSON *et al.*, 2006) onde se tem uma maior velocidade média dos veículos.

Este fator vem sendo associado principalmente às características das vias, como comprimento médio dos trechos, a largura das vias e a hierarquia das mesmas, na medida em que as vias de hierarquia superior tendem a ligar locais da cidade mais distantes ou com outras cidades, sendo até mesmo mais segregadas das demais vias, no caso das estruturais.

### 2.7.4 Características das vias

Conforme visto no capítulo anterior, o fator viário se configura como um dos fatores causadores de acidentes de trânsito. Nesse sentido, diversos autores vêm utilizando os SIG para a obtenção das características geométricas das vias, tais como extensão e largura das mesmas e o número de interseções. UKKUSURI *et al.* (2012) consideram a extensão das vias segmentado pelas diversas larguras de vias, classificação das vias e número de faixas de rolamento.

HA e THILL (2011) classificaram as interseções segundo as vias que as compõem, encontrando mais acidentes envolvendo adultos nos encontros das vias de maior hierarquia predominantemente em regiões comerciais, e maior quantidade de acidentes envolvendo jovens em regiões residenciais, cuja hierarquia das vias das interseções eram menores.

WANG *et al.* (2013) fazem uma revisão bibliográfica sobre o impacto nos acidentes das características do tráfego, tais como velocidade, fluxo, congestionamento e densidade do tráfego, bem como das características de construção das vias.

A análise das características de uma rede de ruas pode revelar importantes relações desta com a quantidade de acidentes. Uma das características mais exploradas é a da conectividade. A conectividade está muito associada à capacidade de se andar a pé em uma dada região, conhecida em inglês por *walkability*. No entanto, o aumento de

conectividade não implica necessariamente no aumento da quantidade de pedestres. JOHNSON e LU (2011) fizeram um estudo sobre a exposição ao risco de crianças e verificaram que a perda de conectividade fazia aumentar o risco de danos graves às mesmas como ocupantes de veículos. MIRANDA-MORENO *et al.* (2011), por exemplo, empregaram as medições de número de interseções, número de trechos de vias e largura média das vias em torno das interseções no estudo dos acidentes envolvendo pedestres em interseções semaforizadas. Segundo ZHANG *et al.* (2013), o impacto da conectividade na segurança de pedestres continua sendo motivo de debate, pois ao mesmo tempo em que uma maior conectividade das vias de uma região tende a aumentar o fluxo de veículos e de pedestres na mesma, por outro lado a velocidade dos veículos tende a diminuir nestas regiões. A conectividade pode ser obtida a partir de diversas medições envolvendo relações entre a área dos quarteirões, extensão das vias e quantidade de interseções. A Tabela 2 apresenta alguns exemplos de medição de conectividade propostos por ZHANG *et al.* (2013), que dividem a conectividade nas categorias métricas, topológicas e comportamentais. A conectividade pode ser representada, por exemplo, pela relação entre o número de trechos de vias e o de interseções.

Tabela 2 Exemplos de medições de conectividade

| **Categoria** | **Medição** | **Definição** |
|---|---|---|
| Métricas | Densidade de interseções | Quantidade de interseções por unidade de área |
| | Densidade de ruas | Extensão das vias por unidade de área |
| | Tamanho do quarteirão | Medida da área ou perímetro do quarteirão |
| Topológicas | Taxa de nós conectados | Quantidade de interseções pelo número de nós* |
| | Relação nó-trecho de via | Quantidade de trechos de vias pelo número de nós |
| Comportamental | Distância a pé | Máximo, médio e mínima distância de casa aos potenciais destinos em uma região |

(*) Considerando os nós como sendo de dois tipos: interseção e final de via.

MARSHALL e GARRICK (2010) mencionam também a análise do padrão da rede de ruas como outro tipo de análise de rede. A Figura 2 apresenta uma adaptação feita por MARSHALL e GARRICK (2011) do conceito de ruas proposto por Stephen Marshall. A partir da observação desta figura é possível ver os padrões das vias da cidade como sendo linear, em árvore ou em grade, sendo que padrão em árvore pode ser tributária ou radial.

O padrão de ruas na vizinhança, por sua vez, pode ser em grade ou em árvore. Da combinação de ambas se teria oito tipos diferentes de padrões de ruas.

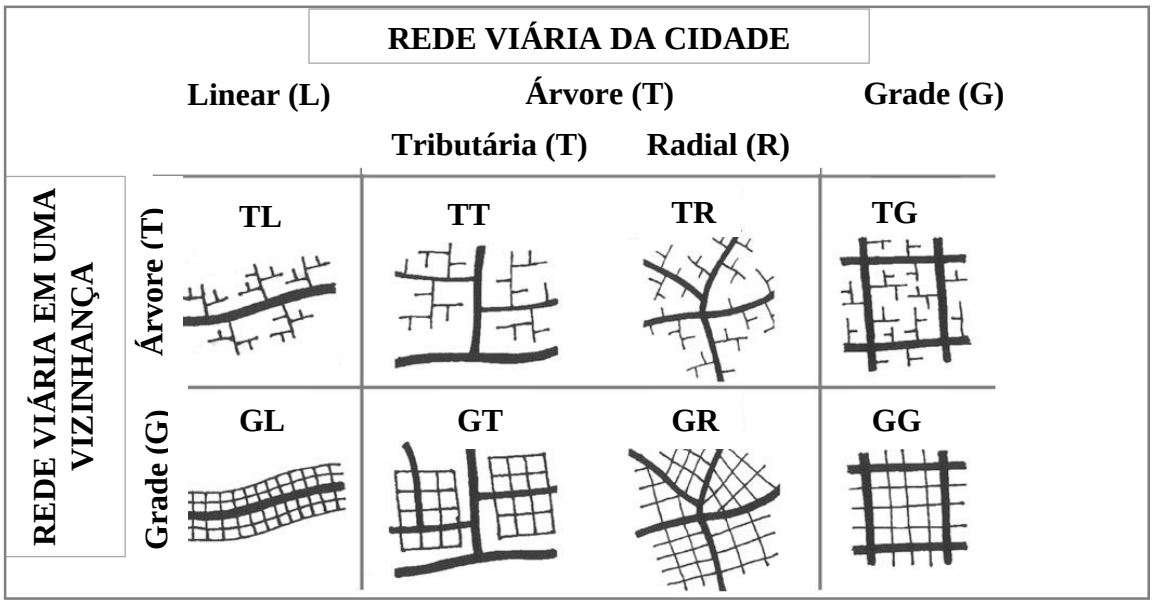


Figura 2 Classificação do padrão de ruas
Fonte: MARSHALL (2005, *apud* MARSHALL e GARRICK, 2011)

Algumas medições de conectividade podem dar uma indicação do padrão de ruas de uma dada região. Como exemplo, a relação entre o número de interseções em X e o número total de interseções pode ser um indicador do padrão de ruas em uma vizinhança ser mais próximo do padrão de uma grade ou de uma árvore. Quanto mais próximos de 1, maior seria a tendência a se ter uma padrão de uma grade. A medição da relação entre o número de interseções em Y e o número total de interseções mais próximos de 1 traria uma maior tendência ao padrão de árvore. A importância da análise do padrão de ruas pode ajudar a entender o porquê da concentração dos acidentes em um maior ou menor número de ruas de uma região. O padrão de grade, por ter maior conectividade, tende a distribuir os acidentes em uma maior quantidade de ruas, principalmente se o valor da taxa de nós conectados for alto, o que indica que as ruas podem ser utilizadas para ligar maiores regiões da cidade e não somente uma vizinhança.

### 2.7.5 Características demográficas e socioeconômicas

Quanto aos fatores demográficos mais explorados na modelagem de acidentes estão a densidade populacional e a faixa etária (NOLAND e QUDDUS, 2004, PULUGURTHA e SAMBHARA, 2011, HA e THILL, 2011, UKKUSURI *et al.* 2012). GRAHAM e

GLAISTER (2003) elaboraram dois índices de atividade de viagens entre bairros como variável indicadora no tráfego entre quarteirões: um em função do nível de emprego e o outro em função da população residente nestes locais.

Quanto aos fatores socioeconômicos, vêm-se empregando a idade, o nível de renda e total de pessoas empregadas como importantes indicadores nesse sentido (NOLAND e QUDDUS, 2004). EDWARDS *et al.* (2006) mostraram que as taxas de acidentes fatais envolvendo crianças cujos pais estavam desempregados ou nunca foram empregados foram mais de 20 vezes maior do que daquelas cujos pais tinham alguma ocupação profissional. COTTRILL e THAKURIAH (2010) encontraram correlação entre os acidentes envolvendo pedestres e o total de crimes em um determinado setor censitário situado em regiões mais carentes.

Alguns autores vêm utilizando para o caso de dados de localidades dos Estados Unidos, o percentual da população pertencente a diferentes raças, principalmente negros, bem como o percentual de imigrantes, com destaque para os de origem hispânica. Um relatório do *Surface Transportation Policy Project* (STPP, 2002) mostrou que em todos os Estados Unidos, para uma população que na época era composta por cerca de 70% de brancos de origem não hispânica, 12% por negros e 12,5% por latinos, obteve um valor de pedestres mortos em acidentes correspondente a 60%, 20% e 13,5% respectivamente.

Estudos envolvendo áreas consideradas carentes foram também estudadas por diversos autores, mostrando comumente maior quantidade de acidentes em regiões mais carentes que nas menos carentes. Nesse sentido, NOLAND e QUDDUS (2004) empregaram o chamado *Index of Multiple Deprivation* (IMD) um indicador socioeconômico inglês na busca de correlação deste com o número e gravidade dos acidentes. LICAJ *et al.* (2011) e COTTRILL e THAKURIAH (2010) trabalharam não com índices, mas com regiões mais carentes. No primeiro caso, as regiões foram as chamadas *Zones Urbaines Sensibles* (ZUS) e no segundo de *Environmental Justice* (EJ).

Diversos estudos também foram feitos mais especificamente sobre o risco das crianças sofrerem acidentes de trânsito (GREEN *et al.*, 2011). DISSANAYAKE *et al.* (2009) encontraram mais acidentes envolvendo crianças em regiões menos densas populacionalmente, principalmente as residenciais e mais próximas de escolas, com mais comércio e menor densidade de interseções. LASCALA *et al.* (2004) encontraram mais colisões envolvendo crianças como pedestres em regiões com maior densidade de população jovem, mais desempregados, menor renda e maior tráfego de veículos.

### 2.7.6 Uso do solo

O uso do solo, por sua vez, é um fator largamente empregado como indicador do fluxo de pedestres e de veículos, gerados pela oferta de habitação, de empregos, de serviços ou de lazer. Nesse sentido, o planejamento urbano também exerce uma forte influência na ocupação do solo, tendo forte efeito na definição do tipo de uso do solo, tais como comercial, residencial e industrial, na extensão das áreas verdes e na densidade de pessoas que ocupam uma determinada área para os mais diversos fins. O tipo de uso do solo vem sendo comumente quantificado nas pesquisas em termos de relação entre o valor da área de cada tipo e o da região com um todo.

CERVERO e RADISCH (1996) mostraram que os residentes de locais mais compactos e com uso do solo mais variado andavam cerca de três vezes mais a pé para acessar as lojas, restaurantes e parques do que aqueles que moram em locais mais espaçados mais propensos ao uso do automóvel. O FHWA (1999) menciona como áreas críticas para acidentes envolvendo pedestres, os centros de atividades urbanas tais como escolas, parques, centros de compras e oferta de transporte público.

Nesse sentido, UKKUSURI *et al.* (2012) encontraram maior correlação entre os acidentes envolvendo pedestres em área comerciais, industriais e áreas abertas do que nas regiões residenciais.

Dentre as atividades consideradas das mais relevantes, a existência de escolas na região de estudo é uma das mais exploradas (CLIFTON e KREAMER-FULTS, 2007, UKKUSURI, 2012). Alguns estudos procuram dar maior enfoque no volume de veículos como é o caso de KINGHAM *et al.* (2011) que procuraram comparar o horário de *rush* de escolas com os demais horários de *rush*. Outros autores procuraram enfocar mais nas crianças como pedestres ou ciclistas. PANTER *et al.* (2010) estudaram o ambiente em torno das escolas na Inglaterra a fim de verificar o quão são mais ou menos adequados para a prática de transporte ativo. Verificaram que quase 50% das crianças iam para a escola a pé ou de bicicleta, bem como que, nas regiões mais carentes, esse percentual tende a diminuir e aumentar nas regiões com maior densidade de ruas. GILES-CORTI *et al.* (2011), por sua vez, verificaram que nas ruas com tráfego mais pesado, tende a diminuir a capacidade das crianças irem a pé para a escola.

SCRIBTER *et al.* (1994) estimaram a relação entre os acidentes e a densidade de quatro tipos de vendas de bebidas alcóolicas: restaurantes, bares, lojas de bebidas e mini-

markets. Verificaram que os locais com maior venda de bebidas alcoólicas apresentavam maior quantidade de acidentes.

Diversos autores (THAKURIAH *et al.*, 2012, COTTRILL e THAKURIAH, 2010) vêm utilizando a disponibilidade de transporte público na modelagem de acidentes. COTTRILL e THAKURIAH (2010) apresentam um Índice de Disponibilidade de Transporte Público, índice que mede a acessibilidade dos residentes de uma dada região ao transporte público, levando em consideração a frequência (pessoas por minuto servidas), horas de serviço (número de horas) e a cobertura do serviço (percentual de área do setor censitário coberto pelo serviço). LICAJ *et al.* (2011) analisaram as características das viagens realizadas por jovens entre 10 e 24 anos entre os municípios da região do *Département du Rhône*, França. Neste estudo, observaram o número de usuários de cada modo e as distâncias percorridas em cada modo, comparando o comportamento dos jovens residentes nos municípios mais carentes com aqueles residentes nos mais ricos, encontrando maior uso de veículos por aqueles pertencentes a famílias com maior renda. Identificaram que as maiores discrepâncias nos deslocamentos ocorreram nos finais de semana, férias e opções de lazer do que no dia-a-dia.

### 2.7.7 Variáveis explicativas a serem empregadas na pesquisa

Após uma revisão bibliográfica mais geral apresentando diversos exemplos do emprego das variáveis explicativas, que não aquelas contidas nos boletins policiais na análise estatística dos acidentes de trânsito, neste item será feita uma apresentação dos estudos que ajudaram a embasar a escolha das variáveis explicativas empregadas nesta pesquisa.

As variáveis associadas à geometria e conectividade das vias são aquelas que vêm sendo mais comumente empregadas na análise dos acidentes de trânsito, com destaque para o somatório da extensão das vias (XU *et al*., 2014, MIRANDA-MORENO *et al*. 2011), número de interseções (SIDDIQUI *et al*., 2012, MIRANDA-MORENO *et al*, 2011, WIER *et al*., 2009) e largura das vias (AGUERO-VALVERDE, 2013). A extensão das vias também aparece dividida por vias de diferentes limites de velocidade (XU e HUANG, 2015), de acordo com a largura e número de faixas de rolamento (UKKUSURI *et al.*, 2012) e também pela hierarquia das mesmas (WIER *et al*., 2009, QUDDUS, 2008). A largura também aparece na forma de largura média das vias (MIRANDA-MORENO *et al*., 2012).

Ao lado da extensão das vias, as variáveis demográficas são as mais comuns na análise de acidentes, com destaque para o somatório da população (PULUGURTHA e SAMBHARA, 2011, MIRANDA-MORENO *et al*., 2011, UKKUSURI *et al*., 2012), a idade da população e a renda. Como variação da população verifica-se a densidade demográfica (XU e HUANG, 2015, WANG e KOCHELMAN, 2013). A idade aparece principalmente na forma de percentual da população dentro de determinadas faixas etárias (HA e THILL, 2011, WIER et al., 2009, NOLAND e QUDDUS, 2004, QUDDUS, 2008), sendo que as mais comumente utilizadas são as faixas abaixo de 18 anos e acima dos 60 anos. A renda, por sua vez, costuma aparecer na forma de renda média (XU e HUANG, 2015, PULUGURTHA e SAMBHARA, 2011).

Dentre as variáveis socioeconômicas, o somatório de pessoas empregadas tem sido a mais empregada (PULUGURTHA e SAMBHARA, 2011, WIER *et al*., 2009, NOLAND e QUDDUS, 2004), bem suas variações como quantidade de desempregados (WIER *et al.*, 2009) e densidade de empregos (NOLAND *et al*., 2013; NOLAND e QUDDUS, 2004). A densidade de empregos também aparece estratificada entres os setores da economia, como serviços e comércio (WANG e KOCHELMAN, 2013). Outra variável socioeconômica também explorada é o número de estabelecimentos e a variação da densidade de estabelecimentos (HA e THILL, 2011).

Em se tratando de acessibilidade aos transportes públicos, vem sendo largamente empregado o número de pontos de ônibus e de estações de metrô (MIRANDA-MORENO *et al*., 2011, PULUGURTHA e SAMBHARA, 2011, QUDDUS, 2008). A extensão de vias de uma dada região utilizada por rotas de ônibus também foi explorada (MIRANDA-MORENO *et al*., 2011).

A seguir, serão apresentados os principais modelos estatísticos empregados na bibliografia para a análise de dados de acidentes de trânsito.

## 2.8 Análise estatística dos acidentes de trânsito

ANASTASOPOULOS e MANNERING (2011) diferenciam os modelos de análise estatística de acidentes entre os modelos de frequência de acidentes, os quais contemplam o número de acidentes ocorridos em um trecho de via ou interseção, independente da severidade da lesão, e os modelos que consideram a severidade das lesões ocorridas nos acidentes, comumente associada à lesão mais grave ocorrida entre os envolvidos no acidente.

ANASTASOPOULOS e MANNERING (2011) mencionam ainda que as variáveis explicativas envolvidas na modelagem que fazem referência aos registros dos acidentes, tais como número dos envolvidos nos acidentes, condições do tempo, etc., podem ser utilizados somente nos modelos de severidade de lesões. Segundo SAVOLAINEN *et al.* (2011), os modelos de frequência de acidentes utilizam outras variáveis que não a dos dados de acidentes, tais como geometria das vias, condições de tráfego, etc.

Tendo em vista que os dados de acidentes utilizados nesta pesquisa não estavam especificados por severidade, os modelos de acidentes de trânsito empregados na pesquisa serão os de frequência de acidentes.

LORD e MANNERING (2010) e SAVOLAINEN *et al.* (2011) fizeram uma revisão bibliográfica sobre os principais modelos de frequência e de severidade, respectivamente, indicando para cada modelo o tipo de erro de inferência o qual se propõe corrigir. Dentre os principais fatores que podem causar tais potenciais erros estão a superdispersão, a subdispersão, a existência de grande quantidade de zeros ou valores abaixo da média, a existência de subregistros de acidentes, a omissão de variáveis explicativas importantes, a consideração de uma inadequada relação entre as variáveis explicativas e a dependente, a desconsideração das correlações espaciais e temporais dos acidentes, dentre outros fatores.

### 2.8.1 Análise estatística de acidentes no Brasil

No Brasil, diversos estudos vêm sendo feitos a partir da espacialização dos dados de acidentes de trânsito. BASTOS (2011), por exemplo, estudou a geografia da mortalidade dos acidentes de trânsito no Brasil, a partir da estimação do valor do índice de mortes por quilômetro percorrido pela frota de veículos rodoviários no Brasil, no período de 2004 a 2008.

Vem-se aplicando diversas técnicas de análise espacial na compreensão da distribuição dos acidentes em um dado local, bem como na correlação das possíveis variáveis explicativas com a distribuição espacial dos acidentes. ALVES (2011) em sua dissertação de mestrado, estudou a correlação entre os acidentes de trânsito e o uso do solo, os polos geradores de viagem e a população residente na cidade de Uberlândia. Aplicou diversas técnicas de estatística espacial para verificar os padrões de distribuição espacial e a existência de dependência espacial dos atropelamentos nos municípios do estado de São Paulo, bem como a correlação com outros indicadores. Nesse sentido,

SANTOS e RAIA JUNIOR (2006) e QUEIROZ *et al.* (2004a) apresentaram a técnica de elipse de desvio padrão para analisar a distribuição dos acidentes. SOUZA (2009) empregou diversas técnicas de análise espacial na compreensão dos acidentes da cidade de Manaus. SOARES (2007) analisou a distribuição espacial dos acidentes a partir da análise de redes na cidade de São Carlos, comparando-se com análises obtidas por SANTOS e RAIA JUNIOR (2006). QUEIROZ (2003) também utilizou diversas técnicas de análise de padrões pontuais para a análise de acidentes na cidade de Fortaleza.

Segundo CUNTO *et al.* (2011), um dos primeiros esforços para a modelagem de acidentes de acidentes no Brasil foi na tese de doutorado de CARDOSO (2006), na qual se modelaram os acidentes de trânsito em vias artérias urbanas na cidade de Porto Alegre. MÂNICA (2007), em sua dissertação de mestrado modelou a distribuição de acidentes nas rodovias do Rio Grande do Sul com a aplicação de modelos de regressão. CUNTO *et al.* (2011) modelaram a distribuição de acidentes de trânsito em interseções semaforizadas na cidade de Fortaleza. BOFFO (2011) fez uma revisão bibliográfica dos principais modelos de previsão de acidentes. HOLZ *et al.* (2011) aplicaram a modelagem da distribuição de acidentes de motocicleta na cidade de Porto Alegre com emprego de regressão linear múltipla. SILVA (2011) fez uma aplicação do manual *Highway Safety Manual* lançado em 2010 para as estradas de pista simples do estado de São Paulo. Embora comumente se empregue modelos estatísticos de acidentes, MADALOZO e DYMINSKI (2009) aplicaram modelos de previsão de acidentes em curvas horizontais utilizando redes neurais e comparou com os modelos estatísticos, em rodovias federais do interior do estado do Paraná. ROCHA e NASSI (2012a) fizeram uma modelagem estatística da distribuição espacial dos acidentes na zona Sul do Rio de Janeiro. BARBOSA *et al.* (2014) aplicaram modelos de previsão de acidentes nas interseções das vias das cidades de Fortaleza, Belo Horizonte e Brasília.

É possível perceber que grande parte das publicações brasileiras envolveu a análise exploratória dos acidentes, conforme mencionado anteriormente, ou envolvendo outras análises voltadas para a epidemiologia dos acidentes de trânsito (FERREIRA, 2009, GOMES e MELO, 2006, MAIA e AIDAR, 2010; MALTA *et al.*, 2011; MORAIS NETO *et al.*, 2012, ROCHA e NASSI, 2012b). Aqueles que trataram sobre modelagem envolveram principalmente dissertações de mestrado e teses de doutorado, sendo que poucos artigos vêm sendo produzidos fora deste escopo.

### 2.8.2 Modelagem de acidentes de trânsito agregados em área

Este item fará uma revisão bibliográfica dos principais modelos estatísticos de acidentes de trânsito utilizados nos últimos anos para o caso dos dados agregados em área. Tendo em vista que este trabalho foca na característica espacial dos dados de acidentes, os modelos de previsão de acidentes serão aqui divididos em dois grupos: os modelos não espaciais e os modelos espaciais. Dentro de cada um dos grupos, por sua vez, os modelos serão apresentados segundo uma ordem crescente de complexidade.

#### 2.8.2.1 Modelos não espaciais de previsão de acidentes

Dentre os modelos estatísticos não-espaciais, os mais simples empregados na modelagem de acidentes de trânsito são os modelos de regressão múltipla, sendo que alguns autores aplicaram transformações nas variáveis do modelo, tais como aquelas a partir da aplicação de uma função logarítmica ou de uma transformação de Box e Cox.

Os MLG com distribuição de Poisson e binomial negativa foram intensamente utilizados nos últimos anos por serem adequados aos dados de contagem, como é o caso dos acidentes de trânsito, sendo que a distribuição binomial negativa contempla o efeito da superdispersão. Mais recentemente, vem-se empregando também os MLG com distribuição de Poisson log-normal. No entanto, segundo HUANG e ABDEL-ATY (2010), os MLG não têm sido capazes de estruturar essa grande dispersão comumente observada nos dados de acidentes. Possuem também a limitação de adotar o pressuposto da independência dos resíduos, desconsiderando dessa forma a possibilidade de haver correlação e heterogeneidade espacial ou temporal dos dados.

Há alguns anos, vem-se também empregando os modelos de Poisson, binomial negativa e Poisson log-normal com a abordagem bayesiana. HUANG e ABDEL-ATY (2010) apresentam diversas vantagens no emprego dos modelos de previsão de acidentes com abordagem bayesiana em relação aos demais modelos.

Alguns autores vêm também empregando os chamados modelos multiníveis, os quais procuram analisar a correlação entre as variáveis em diversos níveis e de forma multitemporal, desde um nível mais macroscópico, que inclui os níveis de região geográfica (área de agregação), local do tráfego (interseções e trechos de vias) e local dos acidentes (severidade dos acidentes, tipo de colisão, etc); até o nível microscópico, que

inclui o nível de unidade motorista-veículo (comportamento do ocupante, manobra do veículo, etc.) e dos ocupantes (sexo, idade, etc.)( HUANG e ABDEL-ATY, 2010, AGUERO-VALVERDE e JOVANIS, 2010, YANNIS *et al.*, 2008). A Tabela 3 apresenta um sumário de alguns artigos utilizados na modelagem não-espacial de acidentes de trânsito agregados em área.

Tabela 3 Sumário dos modelos não espaciais de dados agregados em área

| **Modelos/distribuições** | **Agregação** | **Autores** |
|---|---|---|
| Modelo bayesiano Poisson log-normal e modelo espacial CAR | Agregações de setor censitários | XU *et al.* (2014) |
| Modelos bayesianos completos com binomial negativa, Poisson com ligação log-normal, Poisson inflacionado de zeros, Poisson inflacionado de zeros com ligação log-normal e gama | Trechos de vias | AGUERO-VALVERDE (2013)* |
| Poisson com ligação log, binomial negativa com ligação log e log-normal | Zonas de tráfego | PULUGURTHA *et al.* (2013) |
| Modelo de regressão múltipla com transformação de Box e Cox e binomial negativa | Zonas de tráfego | ROCHA e NASSI (2012a) |
| Binomial negativa e binomial negativa generalizada | Setor censitário e zona de tráfego | UKKUSURI *et al.* (2012) |
| Binomial negativa | *Buffer* em torno de interseções semaforizadas | PULUGURTHA e SAMBHARA (2011) |
| Modelo log-linear, binomial negativa padrão e com ligação log-linear | *Buffer* em torno de interseções | MIRANDA-MORENO *et al.* (2011) |
| Binomial negativa | LSOA (*Lower Super Output Area*) | GREEN *et al.* (2011) |
| Binomial negativa | Setores censitários | MARSHALL e GARRICK (2011) |
| Poisson e Poisson considerando subnotificação exógena | Areas EJ (*Environmental Justice*) | COTTRILL e THAKURIAH (2010) |
| Poisson e binomial negativa | Quadras (*wards*) | DISSANAYAKE *et al.* (2009) |
| Modelos de regressão múltipla com transformação log | Setores censitários | WIER *et al*. (2009) |
| Modelo de regressão múltipla | *Buffer* em torno de interseções | CLIFTON e KREAMER-FULTS (2007) |
| Modelos de regressão múltipla com transformação log | Zonas de tráfego | WASHIGTON *et al.* (2006) |
| Binomial negativa | Células de 0,1 milhas quadradas | KIM e YAMASHITA (2006) |
| Binomial negativa | Quadras (*wards*) | GRAHAM *et al.* (2005) |
| Binomial negativa | Quadras (*wards*) | NOLAND e QUDDUS (2004) |

(*) Utilizado pela diversidade de modelos de previsão de acidentes, embora não seja um modelo de áreas.

#### 2.8.2.2 Modelos espaciais de previsão de acidentes

Conforme mencionado no item anterior, os MLG não são capazes de contemplar a dependência e a heterogeneidade espaciais dos dados. De forma a superar tal limitação, vem-se empregando nos últimos anos os modelos espaciais, tanto na abordagem clássica como na bayesiana.

Dentre os modelos que contemplam a dependência espacial, os mais empregados na análise de acidentes vêm sendo os modelos *Spatial Autoregressive* – SAR (HA e THILL, 2011, QUDDUS *et al.*, 2008) e os modelos *Conditional Autoregressive* – CAR (XU *et al.*, 2014, WANG e KOCHELMAN, 2013, NOLAND *et al.*, 2013, KUHLMANN *et al.*, 2009).

Quanto aos modelos espaciais que contemplam a heterogeneidade espacial, os modelos de regressão ponderada geograficamente (XU e HUANG, 2015, FOTHERINGHAN *et al.*, 2000) vem sendo empregados na previsão de acidentes nos últimos anos (LI *et al.*, 2013, ZHANG *et al.*, 2013).

Tabela 4 Sumário dos modelos espaciais de dados agregados em área

| **Modelos/distribuições** | **Agregação** | **Autores** |
|---|---|---|
| Binomial negativa, binomial negativa com CAR, b*inomial negativa* com parâmetros aleatórios, regressão ponderada geograficamente | Zonas de tráfego | XU e HUANG (2015) |
| Poisson log-normal e Poisson CAR | Setores censitários | XU *et al.* (2014) |
| Poisson log-normal CAR e Poisson log-normal | Polígonos de *Thiessen* a partir de setores censitários | WANG e KOCHELMAN (2013) |
| Binomial negativa com CAR | Setores censitários | NOLAND *et al.* (2013) |
| Poisson log-normal CAR e Poisson log-normal | Zonas de tráfego | SIDDIQUI e ABDEL-ATY (2012) |
| MLG com distribuição binomial negativa, modelos espaciais SAR e CAR bayesiano | Quadras (wards) | QUDDUS (2008) |
| Binomial negativa e modelos bayesianos espaciais (CAR) e espaço-temporais | Condados | AGUERO-VALVERDE e JOVANIS (2006) |
| Regressão ponderada geograficamente com Poisson | Condados | LI *et al.* (2013) |
| Regressão ponderada geograficamente com Poisson | Setores censitários | ZHANG *et al.* (2013) |
| Modelo de regressão múltipla com correção de White e SAR | Setores censitários | HA e THILL (2011) |

WANG *et al.* (2011) fazem uma revisão bibliográfica e uma avaliação metodológica dos modelos espaciais, destacando não somente os modelos que contemplam a dependência e heterogeneidade espaciais, mas também a necessidade de se levar em consideração a questão do MAUP, falácia ecológica e efeito das bordas. Estes dois primeiros foram contemplados por XU *et al.* (2014) e UKKUSURI *et al.* (2012) e o último por SIDDIQUI e ABDEL-ATY (2012).

A Tabela 4 apresenta um sumário de alguns dos artigos que vem sendo empregados na modelagem espacial de acidentes de trânsito agregados em área.

A partir da observação das Tabelas 3 e 4, é possível verificar que a diversidade de áreas de agregação (setores censitários, zonas de tráfego, condados, quadras, *buffer zones* em torno das interseções, etc). É possível também verificar que os modelos espaciais CAR vêm sendo muito mais empregados que os modelos SAR.

Os modelos estatísticos a serem empregados na pesquisa serão melhor explicados posteriormente, no capítulo referente à análise espacial dos dados de acidentes.

# 3. ANÁLISE ESPACIAL

A partir do momento em se associam os dados de acidentes e as demais variáveis utilizadas na pesquisa a uma dada feição espacial, neste caso áreas de agregação, estes dados tornam-se dados geográficos. A seguir, serão comentadas as principais características destes dados, bem como as técnicas de análise espacial que serão empregadas no estudo.

## 3.1 Características dos dados geográficos

Tendo em vista que todas as feições espaciais empregadas no estudo apresentam coordenadas referidas a um sistema geodésico, os dados aqui empregados são considerados dados geográficos, que se diferenciam dos dados espaciais por estarem associados à superfície terrestre. Apresentam três componentes: espacial, não espacial e temporal. A componente espacial está representada em um arquivo gráfico dentro de uma determinada estrutura de dados geográficos, sendo as mais conhecidas a matricial e a vetorial. Cada uma das feições do arquivo gráfico costuma estar associada a registros em uma tabela denominada de tabela de atributos. As componentes não espacial e temporal, por sua vez, estão contidas nas tabelas de atributos.

Partindo da Primeira Lei da Geografia (TOBLER, 1970), que diz que "tudo está relacionado com tudo, mas as coisas mais próximas estão mais relacionadas que as distantes", pode-se chegar a uma característica do dado geográfico que é a da dependência espacial. Outra característica que acompanha os fenômenos geográficos é a da heterogeneidade espacial, chamada por GOODCHILD (2004) de Segunda Lei da Geografia. A heterogeneidade espacial de um dado geográfico faz com que ocorram comportamentos diferentes da variável em locais diferentes da região de trabalho, gerando instabilidade nos coeficientes dos modelos (instabilidade estrutural) e a violação da hipótese de que os erros do modelo sejam constantes ao longo da região (heterocedasticidade).

Em se tratando de dados agregados em área, existem quatro aspectos a serem ressaltados: a questão do MAUP, a falácia ecológica, a falácia atômica e os efeitos de bordas.

O problema conhecido como MAUP – do inglês *Modifiable Areal Unit Problem* (OPENSHAW,1984), diz respeito a existência de instabilidade nos resultados obtidos nas

análises estatísticas dos dados agregados em área à medida em que se muda o nível de agregação do mesmo. Segundo FOTHERINGHAM *et al.* (2000), o MAUP possui dois componentes: o efeito da escala e o efeito da agregação. O primeiro diz que resultados diferentes podem ser obtidos quando aplicada a mesma análise estatística em diferentes níveis de resolução espacial. O segundo afirma que os resultados diferentes de uma mesma análise estatística podem ser derivados de agrupamentos diferentes dos dados dentro de uma mesma escala. Segundo FOTHERINGHAM *et al.* (2000), a solução ideal seria a de utilizar dados desagregados. Na impossibilidade de se ter dados desagregados, a alternativa seria empregar os dados da forma mais desagregada possível e demostrar visualmente os efeitos da mudança de escala e de agregação.

A falácia ecológica decorre do fato de se poder concluir erradamente algo sobre um determinado indivíduo de uma população baseando-se em dados de área. A falácia atomística seria contrária à ecológica, afirmando que se pode fazer uma inferência incorreta para uma agregação baseando-se nos dados no nível dos indivíduos.

OPENSHAW (1984, *apud* WANG *et al.*, 2011) diz que, ao se agregar as variáveis em regiões mais homogêneas, tanto o MAUP como a falácia ecológica tendem a diminuir.

Alguns estudos envolvendo acidentes de trânsito vêm contemplando tais questões, como é o caso de UKKUSURI *et al.* (2012), os quais analisam o efeito da escala nos dados de frequência de acidentes envolvendo pedestres a partir da agregação dos dados de área nos níveis de setores censitários e do código postal americano (*zip codes),* chegando à conclusão que os dados agregados em setores censitários trazem maiores descobertas e são mais consistentes que aqueles no nível dos *zip codes*.

XU *et al.* (2014) analisam as estatísticas obtidas nos níveis de agregação de 734 zonas de tráfego, empregando métodos de regionalização para agrupar os setores censitários de 14 formas diferentes, começando pela divisão da região em 50 partes, e crescendo em intervalos de 50 até atingir o valor de 650 polígonos na região. Verificam que a dependência espacial dos resíduos tende a aumentar à medida em se aumenta a quantidade de áreas de agregação, sendo mais significativos quando se tem um número de áreas maior que 100. Observam também, pela análise da sensibilidade dos coeficientes das variáveis explicativas, que os coeficientes tendem a se estabilizar quando se tem um número de áreas acima de 350 (considerando os intervalos entre 50 e 650 polígonos).

Outro aspecto diz respeito à precisão dos dados, no qual se pode associar uma dada informação a uma área de agregação quando na realidade pertence a outra. Essa questão é particularmente crítica nas bordas da área de agregação. SIDDIQUI e ABDEL-ATY

(2012), após observarem os dados de acidentes em uma dada região, verificaram que grande parte dos mesmos estavam concentrados nas fronteiras das mesmas, o que fez com que propusessem um modelo que levasse em consideração separadamente os acidentes no interior e nas bordas das áreas de agregação.

ALMEIDA (2012) menciona ainda a questão dos *outliers* espaciais como sendo aquelas observações que apresentam dependência espacial distinta das demais regiões e que merecem ser investigados, pois podem ser decorrentes de erros grosseiros de digitação ou podem fornecer importantes informações sobre o fenômeno em estudo.

### 3.2 SIG e análise espacial

Conforme mencionado anteriormente, as variáveis envolvidas na análise espacial tornam-se dados geográficos no momento em que são georreferenciadas e, como tal, poderão ser analisados no contexto de um Sistema de Informações Geográficas (SIG). BURROUGH (1986) define Sistema de Informações Geográficas como sendo "um poderoso conjunto de ferramentas para coletar, armazenar, recuperar, transformar e exibir dados espaciais do mundo real para uma série particular de propósitos". Os componentes de um SIG podem ser mais bem visualizados na Figura 3. O conceito de SIG, por ser multidisciplinar, costuma ter diversas visões dependendo da área do conhecimento na qual está sendo empregado. MAGUIRE *et al*. (1991), por exemplo, menciona três visões, as focadas nos mapas, no banco de dados e na análise espacial. Uma primeira visão estaria mais voltada para o processamento de mapas e está mais ligada às áreas de cartografia e geografia. Em uma segunda estariam os profissionais com maior habilidade computacional onde se destaca as diversas formas de consulta nos bancos de dados e na terceira o SIG seria visto como uma ciência de informação espacial e estaria voltada para análises exploratórias e modelagens espaciais com uso de matemática e estatística. O que diferencia o SIG dos demais sistemas que manipulam os dados geográficos, tais como os de cartografia automatizada e de sensoriamento remoto, é a sua ênfase nas análises espaciais (BURROUGH, 1986, COWEN, 1987 *apud* GOODCHILD, 1987).

HAINING (2003) define análise espacial como sendo uma coleção de técnicas e modelos que empregam explicitamente a posição espacial associada a um dado ou objeto que está especificado dentro do contexto do estudo. CASTIGLIONE (2003) apresenta a análise espacial com enfoque na pessoa que analisa a informação espacial, definindo como sendo a análise de fenômenos que ocorrem no espaço, sendo um processo cognitivo

que só adquire real substância no contexto intelectual de um analista. Todo o processo de análise espacial, que envolve desde a geração de informação geográfica até a interpretação da mesma pelo especialista, pode, inclusive, ocorrer em ambiente não informatizado. Os SIG, com as suas ferramentas computacionais, apresentam grande potencial para suportar complexas operações espaciais e gerar informações significativas. No entanto, para que as informações possam gerar conhecimento é importante que se tenha um analista qualificado que decodifique a informação e gere um bem efetivo.

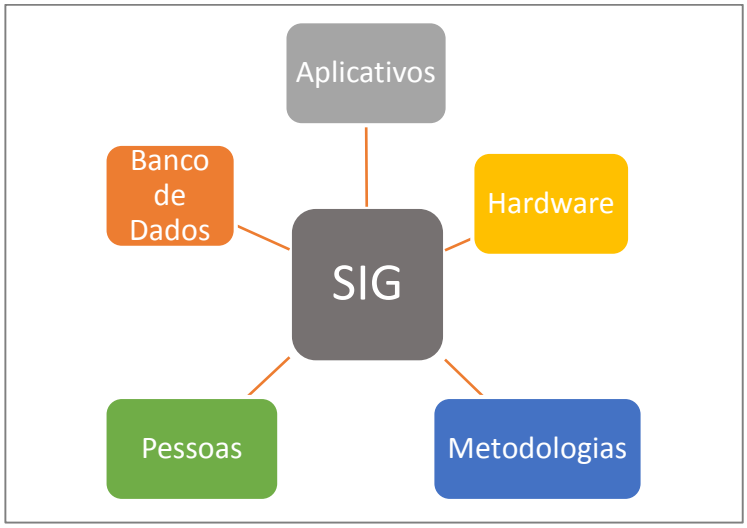


Figura 3 Componentes de um SIG

É importante, portanto, estar atento para não enfatizar excessivamente na tecnologia e esquecer a análise do fenômeno geográfico propriamente dito. CASTIGLIONE (2003) fala sobre o paradoxo que pode ocorrer com o desenvolvimento da tecnologia que diz que quanto mais sofisticados se tornam os instrumentos, maiores são as demandas de conhecimento para lidar com as informações produzidas. Desta forma, quando se busca espacializar um dado fenômeno, é importante conhecer não somente sobre o tema no qual se trata o fenômeno, mas também a ontologia dos processos espaciais para que o entendimento seja compatível com as informações produzidas no SIG. A geografia como uma ciência que analisa o espaço geográfico e a cartografia como ciência que representa os fenômenos geográficos podem oferecer uma importante contribuição nesse sentido.

Alguns autores diferenciam a análise espacial da análise de dados espaciais. HAINING (2003), por exemplo, divide a análise espacial em três elementos: a análise cartográfica, representada por mapas e por operações obtidas a partir de mapas; a análise matemática, onde o modelo obtido é dependente da forma de interação espacial entre os

objetos do modelo e a análise estatística, esta última conhecida como análise de dados espaciais.

BAILEY e GATRELL (1995) definem a análise de dados espaciais como sendo uma parte da análise espacial em que se tem um conjunto de dados observados de um processo espacial e se deseja, por meio da aplicação de diversos métodos, descrever e explicar o comportamento desse processo e sua possível relação com um dado fenômeno espacial.

Como é possível verificar, talvez por questões de especialização, os conceitos de SIG e de análise espacial sejam abordados com uma visão compartimentada entre as áreas de cartografia, computação e matemática/estatística, ou combinadas com outras áreas, como é o caso da economia produzindo a Econometria Espacial. Tal forma de trabalhar embora seja muitas vezes funcional pode gerar uma visão limitada do fenômeno. Esta tese busca abordar o assunto dos acidentes de trânsito da cidade do Rio de Janeiro com uma abordagem integrada entre estas diversas áreas, de modo a buscar melhor analisar o fenômeno em estudo.

Por questões didáticas, a análise espacial será aqui dividida entre análise visual, análise exploratória e modelagem, conforme foi proposto por ANSELIN (2002), embora sabendo que estas três etapas se misturam ao se realizar a análise espacial.

### 3.3 Análise visual dos dados geográficos

A análise visual dos dados geográficos vem passando nos últimos anos por uma mudança no seu paradigma, indo de uma representação estática dos mapas em papel para uma visualização mais sofisticada em ambiente digital. Nesse sentido, MAC EACHREN (1994) já destacava, naquela época, certa encruzilhada da Cartografia entre as tradições do passado e o SIG. PETERSON (1994, *apud* CASTIGLIONE, 2003) destaca a interatividade e a animação como importantes elementos à compreensão da espacialidade dos fenômenos, sendo que os SIG vêm explorando com mais intensidade o primeiro que o segundo elemento.

Nesse contexto, a Cartografia, tradicionalmente preocupada com a comunicação visual de um determinado tema, vem cada vez mais sendo influenciada pela área da computação gráfica denominada de Visualização Científica Computacional. Enquanto que a Cartografia é mais caracterizada por uma representação de domínio público mais que de domínio privado, a Visualização Científica busca no domínio privado e com mais interatividade a exploração dos dados e informações visando à compreensão do

fenômeno. No entanto, considerando a interatividade, o uso privado de um mapa e a busca por incógnitas, MAC EACHREN (1994) afirma que a diferença entre a comunicação e a visualização estaria na combinação destes três fatores, como pode ser visto pelo cubo de MAC EACHREN (1994) constante da Figura 4.

Como se pode perceber, a visualização e a comunicação estão situadas em posições diametralmente opostas. Quando o operador habilitado explora visualmente uma variável ele tende a recorrer a diversas formas de visualização. No entanto quando gera informação espacial para uso público ou acessa essa informação disponibilizada por outrem, o enfoque no uso do mapa está na comunicação da informação.

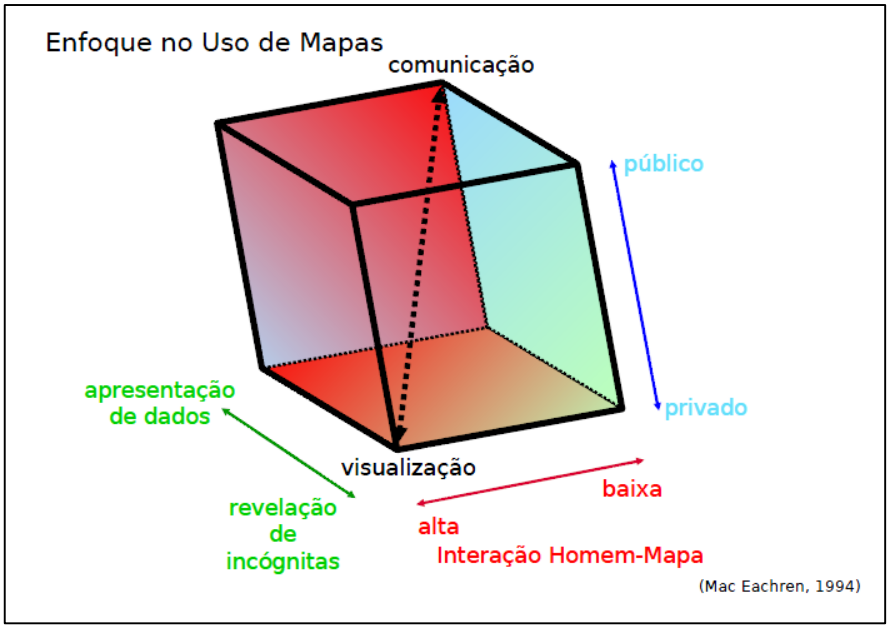


Figura 4 Cubo de Mac Eachren

Fonte: MAC EACHREN (1994)

Uma das formas de representação espacial do dado geográfico é por meio dos mapas temáticos. Os mapas temáticos, como o próprio nome já diz, são mapas empregados para atender a uma determinada finalidade ou tema. Segundo DENT (1985), os mapas temáticos podem ser classificados nos mapas coropléticos, com símbolos proporcionais e densidade de pontos. O tipo de mapa temático que costuma ser mais empregado para os dados de área, os quais serão empregados nesta pesquisa, são os coropléticos, que é um tipo de representação em que se seleciona um número de classes para a legenda e se associa diferentes tons de cores para as classes, de forma a representar a intensidade do fenômeno em estudo (DENT, 1985). Dependendo da quantidade de classes e do critério empregado na construção da legenda, podem-se ter visualizações diferentes da variável

de interesse. As Figuras 5 e 6 apresentam dois mapas temáticos dos acidentes na zona Sul do Rio de Janeiro agregados nas zonas de tráfego para o ano de 2011, um construído pelo critério dos quartis e outros pelo critério da quebra natural. É possível observar diferenças nos valores da legenda e consequentemente da distribuição das cores no mapa. O primeiro agrupa em uma classe da legenda a mesma quantidade de zonas de tráfego e o segundo agrupa em uma classe da legenda as zonas de tráfego cujos valores estão mais próximos entre si, ou seja, com menor variância.

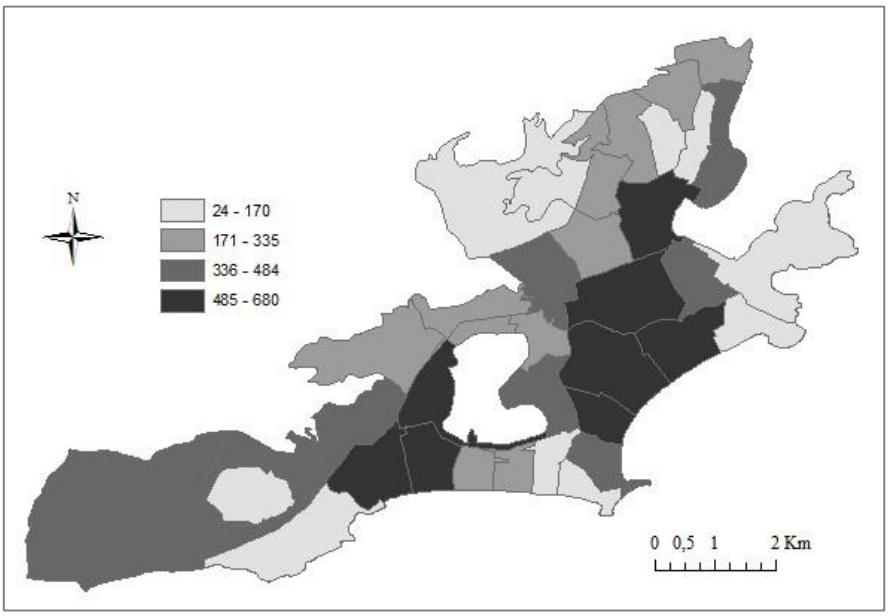


Figura 5 Mapa coroplético dos acidentes produzido pelo critério dos quartis

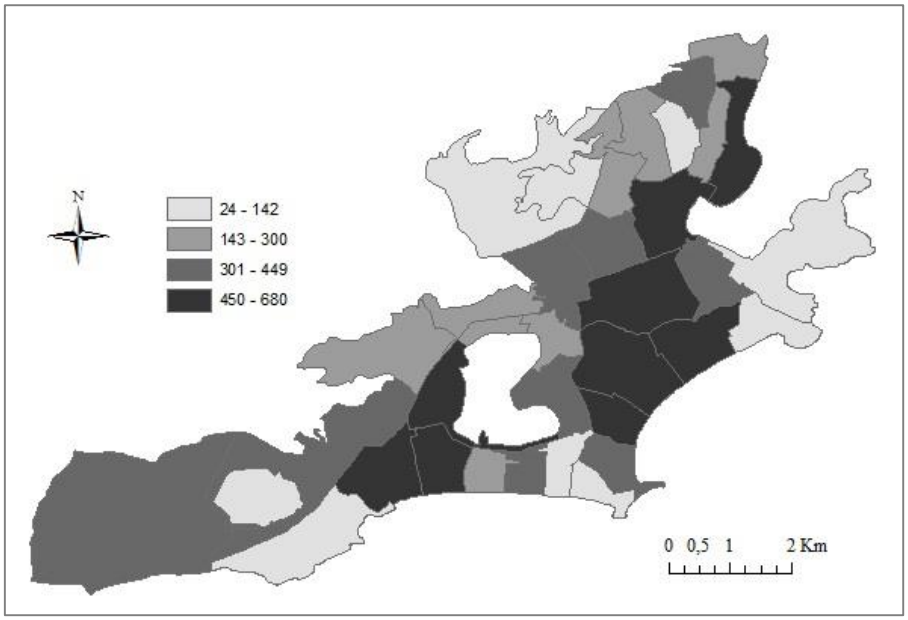


Figura 6 Mapa coroplético dos acidentes pelo critério da quebra natural

A visualização destas feições espaciais pode ser feita de diversas formas. BERTIN (1967) propôs, dentro da Semiologia Gráfica, diversas variações visuais que poderão ser aplicadas na visualização cartográfica. São elas a forma, orientação, cor, valor e granulação, conforme pode visto na Figura 7.

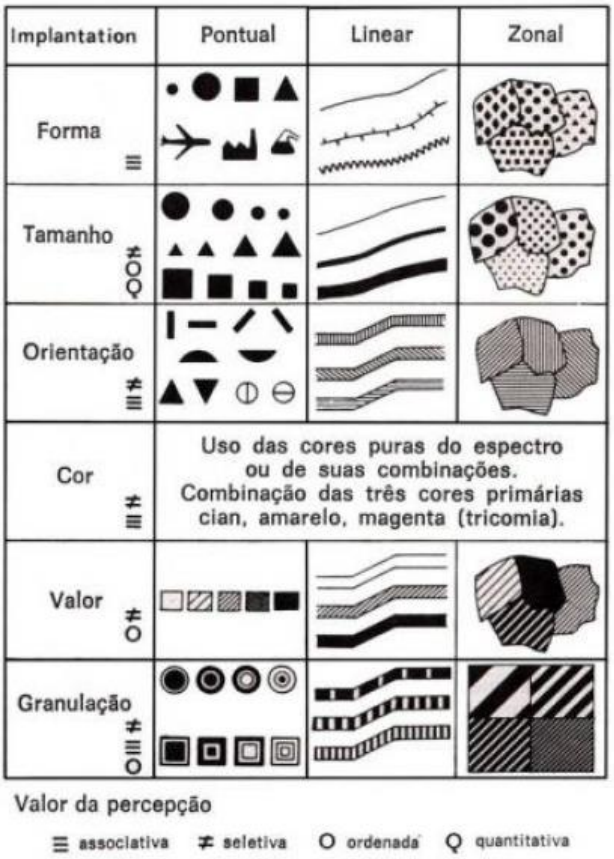


Figura 7 Variáveis visuais propostas por Bertin

Fonte: BERTIN (1967)

A partir da observação dos mapas constantes das Figuras 5 e 6, é possível verificar que os mapas coropléticos apresentam a representação espacial que emprega a variação de cor proposta por Bertin.

Quando se tem a posição do acidente representada espacialmente por uma feição pontual, é possível em muitos casos identificar um padrão de distribuição dos pontos diretamente em um aplicativo de SIG ou mapa sem aplicação de nenhum operador espacial. A Figura 8 apresenta o exemplo das feições pontuais de acidentes na zona Sul do Rio de Janeiro. Outras vezes, podem-se produzir superfícies a partir destes dados, como é o caso da superfície de densidade de Kernel, empregada por alguns autores para a visualização dos dados de acidentes (ROCHA e NASSI, 2012b, HA e THILL, 2011).

A Figura 9 apresenta o exemplo da superfície de Kernel gerada a partir dos dados constantes da Figura 8. Estes dados pontuais também poderão ser associados a feições lineares e de área. No primeiro caso, costuma-se construir em um primeiro momento *buffer zones* em torno das feições lineares para depois fazer a interseção dos dados de acidentes a estas feições. No caso das feições de área, associa-se os pontuais diretamente às mesmas.

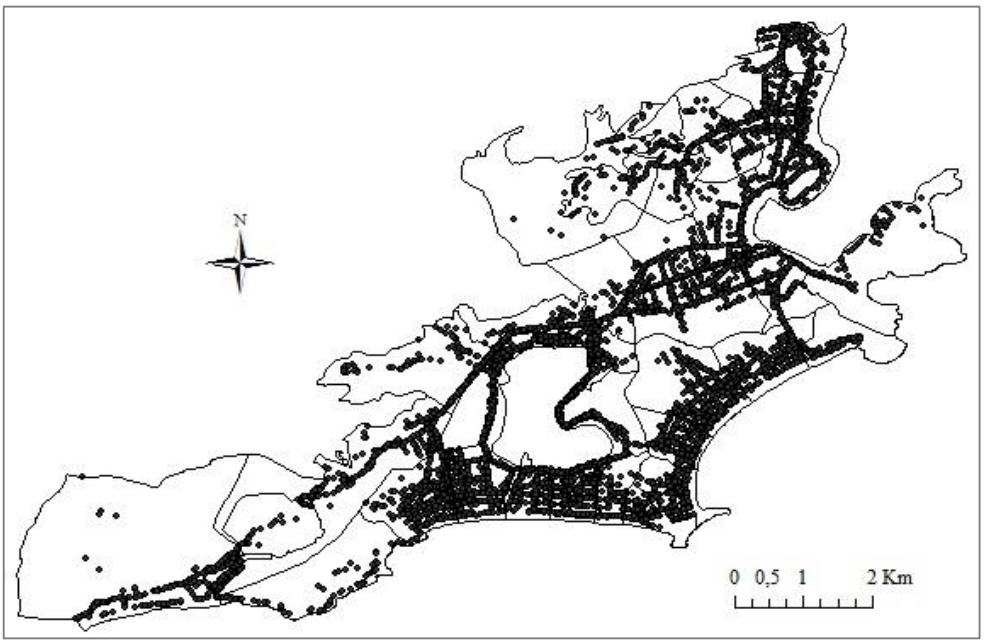


Figura 8 Dados pontuais de acidentes na zona Sul do Rio de Janeiro

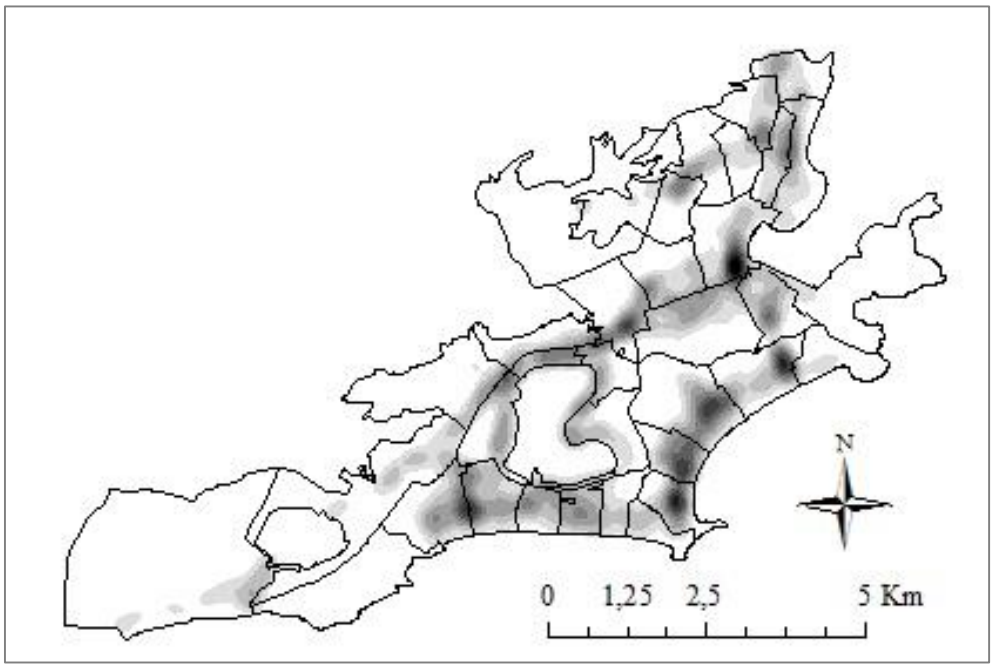


Figura 9 Mapa de Kernel dos bairros da zona Sul do Rio de Janeiro

## 3.4 Análise exploratória dos dados de acidentes

Em uma modelagem estatística, a análise exploratória é a etapa que antecede a modelagem propriamente dita. A análise exploratória de dados é o processo que utiliza medidas estatísticas, tabelas, gráficos e outras representações de um conjunto de dados, com vistas a compreender melhor o conjunto de dados e, desse modo, identificar padrões e relacionamentos nos mesmos. A análise exploratória será aqui dividida em análise exploratória não espacial e análise exploratória espacial.

### 3.4.1 Análise exploratória não espacial

A análise exploratória não espacial é a clássica Estatística Descritiva, que inclui medidas tais como a média, a moda, a mediana e o desvio-padrão, bem como diversos tipos de gráficos, tais como os diagramas de caixas (*boxplots*), histogramas, diagramas de dispersão (*scatter plots*) etc. Tal análise utiliza valores da tabela de atributos do dado geográfico na obtenção das análises.

### 3.4.2 Análise exploratória espacial

Na análise exploratória espacial, algumas técnicas estatísticas são empregadas para identificar padrões de associação espacial e aglomerados (*clusters*), a partir da consideração da distribuição espacial do fenômeno em estudo, com vistas a identificar a existência de dependência espacial e variabilidade da variável em estudo no espaço.

A noção de dependência espacial costuma ser materializada a partir do cálculo da autocorrelação espacial, derivada da autocorrelação empregada na análise de séries temporais que, por sua vez, vem do conceito de correlação. A autocorrelação associa a uma mesma variável em períodos de tempo diferentes. Já a autocorrelação espacial, utiliza a mesma variável em locais diferentes. Quando ocorre dependência espacial, o pressuposto de independência dos resíduos não é obedecido e tem-se que aplicar modelos inferenciais que possam levar em consideração tal efeito. No próximo item serão vistos os principais indicadores de autocorrelação espacial.

A heterogeneidade espacial, por sua vez, pode ser identificada quando se empregam técnicas que possam realçar a existência de comportamentos distintos do fenômeno na

região de estudo, o que pode se refletir na variação dos coeficientes da regressão e no intercepto. Tal comportamento pode ser detectado a partir da construção de mapas temáticos e da comparação dos mesmos com os diagramas de dispersão de Moran, conforme propõe ALMEIDA (2012), pela técnica de ANOVA espacial, superfícies de tendência ou pela utilização de métodos de regionalização espacial, como é caso do proposto por XU *et al.* (2014).

#### 3.4.2.1 Indicadores de autocorrelação espacial

Conforme mencionado anteriormente, a dependência espacial é representada por meio dos indicadores de autocorrelação espacial, os quais podem ser classificados em globais ou locais. Nos globais, busca-se verificar a existência de algum padrão de dependência espacial na variável em toda a área de estudo, podendo ter uma distribuição aleatória, sistemática ou em aglomerados. No caso dos locais, busca-se identificar os padrões na vizinhança de cada área. Dentre os diversos índices de autocorrelação espacial existentes, serão empregados na pesquisa o índice global e local de Moran.

Um conceito muito importante e que irá compor a formulação da autocorrelação é a matriz de proximidade, representada por uma matriz $w_{ij}$ onde o índice *i* está associado a um dado polígono e o índice *j* representa o polígono vizinho de *i*. A definição desta matriz pode empregar diversos critérios. Dentre os mais conhecidos estão o da torre, a qual considera como vizinho de um polígono aqueles que compartilham lados comuns e o da rainha que considera como vizinhos aqueles que compartilham vértices com o polígono *i*. Outros critérios comuns são o de adotar como vizinhos aqueles polígonos situados dentro de uma distância máxima em torno do centroide de *i*, bem como o de adotar os *n* polígonos cujos centroides estão mais próximos do centroide de *i*.

##### 3.4.2.1.1 Índice global de Moran

O índice global de Moran pode ser representado pela Eq. 1.

$$I = \frac{n\,(X_i - \bar{X}) \sum_i \sum_j w_{ij}(X_j - \bar{X})}{\sum_i \sum_j w_{ij} \sum_i (X_i - \bar{X})^2} \qquad \text{Eq. 1}$$

Onde *n* é o número de áreas, $X_i$ é o valor do atributo considerado na área *i*, $\bar{X}$ representa o valor médio do atributo na região de estudo, e $w_{ij}$ os pesos atribuídos conforme a relação entre as áreas *i* e *j*. Sendo um indicador global, determina-se para toda

a região de estudo somente um valor deste índice. Para valores do índice de Moran próximos a 0, tem-se a situação de independência espacial, para aqueles próximos de -1 de regularidade e de +1 a existência de aglomerados ou *clusters*. A distribuição espacial da variável em torno de cada área pode ser mensurada pelo índice local de Moran.

#### 3.4.2.1.2 Índice local de Moran

O índice de local de Moran, também chamado de LISA (*Local Indicators of Spatial Association*) para cada área *i* pode ser obtido pela Eq. 2.

$$I_i = \frac{n\,(X_i - \bar{X}) \sum_j w_{ij}(X_j - \bar{X})}{\sum_j (X_j - \bar{X})^2} \qquad \text{Eq. 2}$$

Onde *n* é o número de áreas, $X_i$ é o valor do atributo considerado na área *i*, $\bar{X}$ representa o valor médio do atributo na região de estudo, e $w_{ij}$ os pesos atribuídos conforme a relação entre as áreas *i* e *j*.

#### 3.4.2.1.3 Diagrama de dispersão de Moran

O diagrama de dispersão de Moran é uma representação gráfica que compara o valor do desvio do valor da variável em relação à média global com o valor da média móvel espacial. O desvio padronizado da média global $z_i$ está representado pela Eq. 3.

$$z_i = \frac{(y_i - y_{med})}{\sigma_y} \qquad \text{Eq. 3}$$

Sendo $y_i$ o valor da variável e $y_{med}$ a média aritmética da variável e $\sigma_y$ o desvio-padrão de *y*.

A média móvel espacial para cada área *i* pode ser representada pela Eq. 4 (BAILEY e GATRELL, 1995).

$$\hat{\mu}_i = \frac{\sum_{j=1}^{n} w_{ij} y_j}{\sum_{j=1}^{n} w_{ij}} \qquad \text{Eq. 4}$$

Onde *i* = 1, 2, ..., *n*, e $y_i$ é o valor do atributo na área *i*, *n* é o número de áreas e $w_{ij}$ é o elemento da matriz de vizinhança que representa as relações existentes entre as áreas.

O valor do desvio padronizado em relação à média móvel seria então representado pela Eq. 5.

$$wz_i = \frac{(\hat{\mu}_i - \hat{\mu}_{i_{med}})}{\sigma_{\hat{\mu}_i}} \qquad \text{Eq. 5}$$

O diagrama de dispersão de Moran apresenta os resultados em quatro classes: Alto-alto, Alto-baixo, Baixo-baixo e Baixo-alto nos quadrantes de I a IV, respectivamente, referindo-se aos valores de $Z_i$ e $WZ_i$, nessa ordem (ANSELIN, 1996), conforme pode ser visto na Figura 10. No eixo das abscissas constam os valores do desvio da média dos valores da variável, sendo que na origem consta o valor da média aritmética da variável. No eixo das ordenadas constam os valores do desvio das médias móveis espaciais, sendo que a origem é o valor médio das médias espaciais.

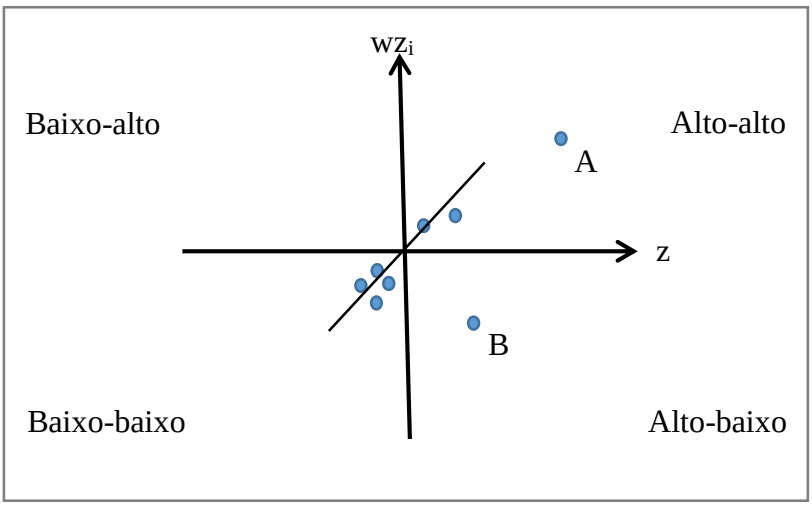


Figura 10 Quadrantes do diagrama de dispersão de Moran

Segundo ALMEIDA (2012), o diagrama de dispersão de Moran podem revelar *outliers* espaciais, ou seja, pontos afastados dos demais pontos do diagrama e que podem afetar os valores da autocorrelação espacial, prejudicando a estimativa do teste.

ALMEIDA (2012) diferencia os *outliers* espaciais dos pontos de alavancagem, sendo estes últimos situados no mesmo quadrante que o conjunto dos demais dados enquanto que os *outliers* espaciais não constam dos mesmos quadrantes da maioria dos dados. O ponto A do diagrama de dispersão de Moran constante na Figura 10 seria um exemplo de ponto de alavancagem enquanto que o B seria um exemplo de *outlier* espacial.

A Figura 11 mostra o diagrama de dispersão de Moran construído a partir dos dados de acidentes apresentados na Figura 8. A Figura 12 apresenta o mapa de Moran construído a partir dos mesmos dados. Cada uma das áreas do mapa representa um ponto no diagrama, sendo que cada uma das classes da legenda representa os pontos de um dado quadrante. Este mapa também é conhecido como mapa LISA.

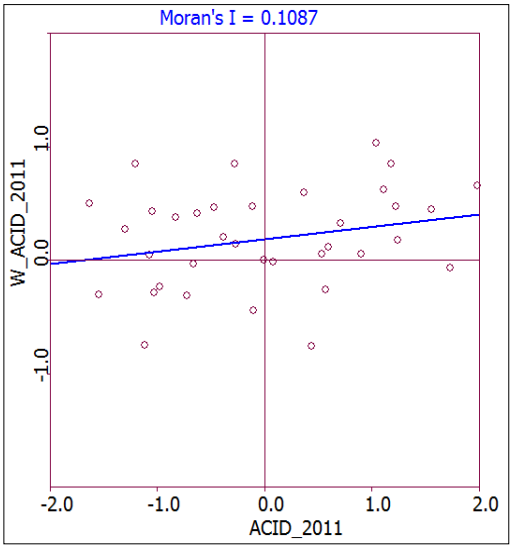


Figura 11 Diagrama de dispersão de Moran dos acidentes em 2011

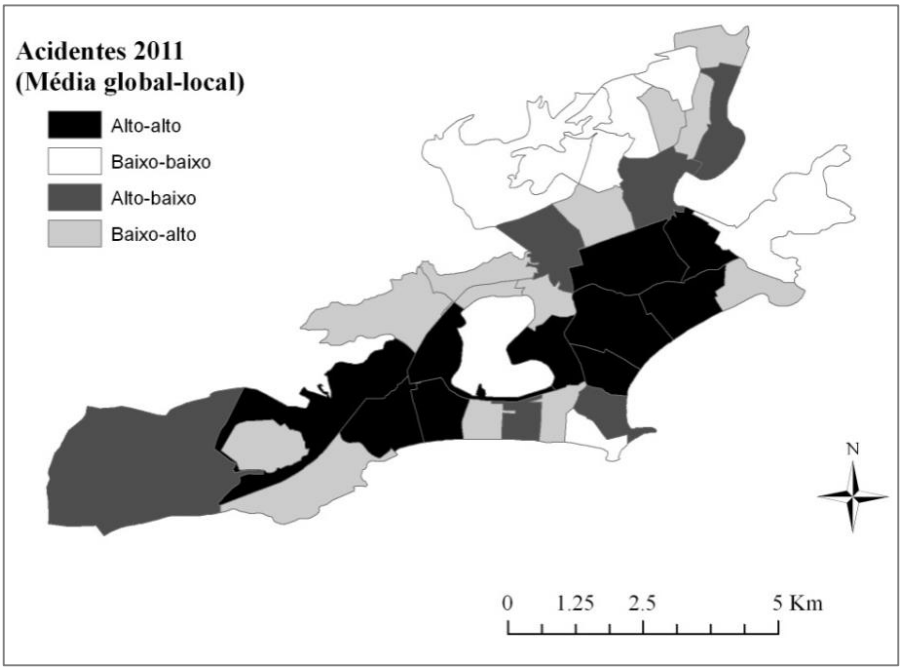


Figura 12 Mapa de Moran dos dados de acidentes na zona Sul do Rio de Janeiro

### 3.5 Modelagem estatística

Os modelos estatísticos a serem empregados na pesquisa estarão divididos entre os modelos não espaciais e os modelos espaciais. De todos os modelos apresentados na revisão bibliográfica, serão apresentados somente aqueles a serem utilizados na metodologia.

### 3.5.1 Modelos não espaciais

Os modelos não espaciais empregados na metodologia serão os modelos de regressão múltipla e os modelos lineares generalizados com distribuição de Poisson e binomial negativa. Estes modelos são considerados não espaciais por utilizarem somente os valores da tabela de atributos dos dados geográficos e por não empregar nenhuma variável que possa de alguma forma refletir algum tipo de relação espacial, como é o caso das matrizes de ponderação.

#### 3.5.1.1 Modelos de regressão múltipla

Um dos modelos não espaciais mais simples empregados na modelagem estatística de acidentes é o modelo de regressão múltipla, representado pela Eq. 6.

$$Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \varepsilon \qquad \text{Eq. 6}$$

Onde *Y* representa a variável resposta, aqui representada pelo número de acidentes em uma dada área de agregação, $X_n$ representam as n diferentes variáveis explicativas, $\beta_t$ representam os *t* coeficientes da equação e a variável $\varepsilon$ representa o resíduo da regressão.

Dentre as condições para a aplicação de um modelo de regressão linear múltipla, estão a de que o resíduo ε tenha uma distribuição normal com média igual a 0 e variância constante (homocedasticidade) e que sejam independentes. Na prática, para verificar a normalidade dos resíduos, costuma-se construir o histograma ou o Normal Q-Q sobre a variável dependente ou aplicar um teste de normalidade sobre a variável dependente tal como Anderson-Darlin e Shapiro-Wilks. Caso não se verifique a condição de normalidade, pode-se aplicar uma transformação sobre a variável dependentes, como é o caso da transformação de Box e Cox (1964) sobre a variável dependente antes de empregar a modelagem, conforme Eq. 7.

$$z = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} \; se \lambda \neq 0 \\ \log(y) \;\; se \lambda = 0 \end{cases} \qquad \text{Eq. 7}$$

Onde λ é uma constante desconhecida. A ideia dessa transformação é encontrar um valor de λ de tal forma que a nova variável se torne aproximadamente normal e, além disso, consiga produzir constância na variância e a linearidade entre a esperança da nova

variável z e as covariáveis selecionadas para compor o modelo.

Outros gráficos são necessários para verificar os pressupostos da regressão linear, conforme apresenta CORDEIRO e DEMÉTRIO (2008). ANSELIN (2005) propõe ainda alguns testes para a verificação dos pressupostos da regressão linear múltipla. Sugere o número de condição de colinearidade que não é um teste, mas um número que serve de diagnóstico de colinearidade e o teste de Jarque-Bera para a normalidade.

#### 3.5.1.2 Modelos lineares generalizados

Os modelos lineares generalizados foram propostos por Nelder e Weddeburn em 1972 como sendo uma extensão dos modelos lineares clássicos, cujas suposições de normalidade, homogeneidade de variâncias e relacionamento linear entre os efeitos da covariável e a média eram relaxados, produzindo uma abordagem unificada para análise de uma ampla classe de dados contínuos e discretos.

Segundo CORDEIRO e DEMÉTRIO (2008), um modelo linear generalizado é definido por uma distribuição de probabilidade para a variável resposta pertencente à família exponencial, formada de uma componente aleatória, a qual especifica a distribuição de probabilidade da variável resposta, uma componente sistemática, a qual especifica uma função linear das variáveis explicativas e uma função de ligação que relaciona uma combinação linear das variáveis explicativas com o valor médio da variável resposta.

A componente aleatória de um MLG considera que se dispõe de um vetor de observações $y=(y_1, ..., y_n)^T$ como realização das variáveis aleatórias $Y=(Y_1, ..., Y_n)^T$ independentes ou pelo menos não correlacionadas, cada uma com distribuição pertencente à família exponencial, que tem as seguintes propriedades:

a. A distribuição de cada $Y_i$ é da forma canônica depende de um único parâmetro, digamos $\theta_i$, onde os $\theta_i$´s não tem que ser os mesmos, sendo expressa pela Eq. 8.

$$f(y_i;\theta_i) = exp[y_i b_i(\theta_i) + c_i(\theta_i) + d_i(y_i)] \qquad \text{Eq. 8}$$

b. A distribuição de todos os $Y_i$´s é da mesma forma, ou seja, a função densidade conjunta de $Y_1, Y_2,..., Y_n$ pode ser escrita conforme Eq. 9.

$$y(y_1, \dots, y_n; \theta_1, \dots, \theta_n) = exp\left[\sum_{i=1}^{n} y_i b_i(\theta_i) + \sum_{i=1}^{n} c_i(\theta_i) + \sum_{i=1}^{n} d_i(y_i)\right] \qquad \text{Eq. 9}$$

Distribuições conhecidas como a binomial, Poisson, gaussiana, gaussiana inversa e Gamma, só para citar alguns exemplos, são membros da família exponencial. Além disso, para a escolha de uma distribuição adequada para o erro, deve-se examinar o tipo de dado, especialmente quanto aos aspectos de assimetria, natureza contínua ou descontínua e intervalo de variação. Essa componente é representada pelo erro aleatório ε no modelo linear clássico.

Essa componente, também chamada de preditor linear, é definida como uma função linear dos parâmetros desconhecidos $\beta=(\beta_i,...,\beta_k)^T$ representado no modelo linear por $\beta X$, onde $X$ é a matriz, de dimensões $n$ x $k$ do modelo. A escolha do preditor linear adequado leva em consideração técnicas de seleção de covariáveis.

Essa função faz a ligação entre a média da variável resposta e a estrutura linear do modelo. Assim, a escolha dessa função deve ser compatível com a distribuição proposta para o erro, de forma a facilitar a interpretação do modelo.

A estimação do vetor de parâmetros desconhecidos é feita utilizando-se o método da máxima verossimilhança, quando o sistema de equações é linear. Caso contrário, necessita-se de métodos alternativos, como o método de Newton-Raphson para se determinar $\widehat{\beta}$, ou seja, a estimativa de $\beta$.

A qualidade do ajuste de um MLG é medida por meio da função desvio (*deviance*) dada pela Eq. 10.

$$S_p = 2\left(\hat{l}_n - \hat{l}_p\right) \qquad \text{Eq. 10}$$

Onde $S_p$ é o desvio do modelo, $\hat{l}_n$ e $\hat{l}_p$ são respectivamente, os máximos da log-verossimilhança para o modelo saturado (número de observações $n$ igual ao número de parâmetros $p$ dos coeficientes $\beta$'s) e para a investigação ($p<n$).

O desvio é uma medida de distância entre os valores ajustados e observados e, segundo CORDEIRO e DEMÉTRIO (2008), embora pouco seja conhecido sobre a distribuição do desvio, na prática compara-se $S_p$ com o valor crítico $\chi^2_{n-p;\alpha}$ da distribuição qui-quadrado a um nível de significância α com $n$-$p$ graus de liberdade.

Os modelos com distribuição de Poisson podem ser representados pela forma log-linear pela Eq, 11.

$$\ln(\lambda_i) = \beta_0 + X_i\beta + \theta_i \qquad \text{Eq. 11}$$

Onde $Y_i \sim$ Poisson ($\lambda_i$).

Onde $\lambda_i$ é igual ao valor esperado $E[y_i]$ de $y_i$ e $var(y_i) = E[y_i]$. Onde $\theta_i$ é um erro aleatório não correlacionado com X. Estes erros, por sua vez, podem ser introduzidos seja

por variáveis não observadas no modelo seja por uma fonte pura de aleatoriedade (NOLAND e QUDDUS, 2004). Dessa forma, $\lambda_i$ varia em função de X é tal variação é inserida pela heterogeneidade observada.

A distribuição binomial negativa, tem-se a seguinte forma log-linear dada pela Eq. 12.

$$\ln(\lambda_i) = \beta_0 + X_i\beta + \theta_i \qquad \text{Eq. 12}$$

Onde $Y_i \sim$ Binomial negativa ($\lambda_i$, $\alpha$), $\lambda_i$ é o valor esperado de $y_i$ e var[$y_i$] = E[$y_i$] + $\alpha$E[$y_i$]$^2$. Onde $\theta_i$é um erro com distribuição Gamma com média 0 e variância α. aleatório não correlacionado com X. Neste caso, $\lambda_i$ varia em função não somente de X, mas também por uma heterogeneidade não observada no erro $\theta_i$ (NOLAND e QUDDUS, 2004).

### 3.5.2 Testes de dependência espacial

Os modelos estatísticos não espaciais desconsideram os chamados efeitos espaciais, que são a dependência espacial e a heterogeneidade espacial. Os modelos espaciais são empregados quando é verificado que o valor de uma variável em um local *i* pode ser dependente do valor da mesma variável nos vizinhos *j*, de outras variáveis nos vizinhos *j* ou dos erros da modelagem, estes últimos podendo ser causados pela não consideração de alguma variável explicativa que contemple a dependência espacial da variável resposta.

Segundo ALMEIDA (2012), existem três possíveis causas de dependência espacial: a interação espacial, o erro de medida dos dados espaciais e a má especificação do modelo. A interação espacial diz respeito ao fato de que eventos em um dado local podem afetar não somente este local, mas também outros locais. O erro de medida ocorre quando existe uma diferença entre o nível de agregação em que ocorre o fenômeno e aquele que se encontram as variáveis explicativas. A má especificação do modelo ocorre quando alguma variável explicativa relevante que capte a dependência espacial é omitida do modelo. Poderia ainda acrescentar a influência dos *outliers* e pontos de alavancagem vistos anteriormente.

Segundo ALMEIDA (2012), a heterogeneidade espacial pode ser causada pelo erro de medida dos dados espaciais, pela má especificação do modelo e pelas diferenças de estrutura espacial do modelo de acordo com a região de trabalho. Os dois primeiros itens foram tratados anteriormente. As diferenças de estrutura espacial podem fazer com que

ocorram, por exemplo, diferentes regimes espaciais na área de estudo, fazendo com que os coeficientes apresentem valores diferentes, na medida em que se percorrem diferentes regiões da área de trabalho, levando a diferentes valores da variância nestas regiões (heterocedasticidade).

Pode também ocorrer que um dos efeitos espaciais possa estar causando o outro efeito. Como exemplo, pode ocorrer dependência espacial devido à desconsideração do comportamento que o modelo possa ter em diferentes regiões da área de estudo. Por sua vez, a existência de dependência espacial não adequadamente contemplada no modelo, pode levar à heterogeneidade espacial. Como forma de atenuar tal situação, recomenda-se as análises visual e exploratória, de forma a melhor compreender a distribuição espacial do fenômeno em estudo e tentar diminuir os erros de medida, bem como identificar possíveis regimes espaciais e interações entre as variáveis.

Antes de serem aplicados os modelos espaciais, é preciso verificar primeiramente a existência de dependência espacial para depois verificar a existência de heterogeneidade espacial. Como forma de detectar a autocorrelação espacial e, dessa forma, auxiliar na especificação e validação dos modelos espaciais, pode-se aplicar alguns testes. Segundo ANSELIN *et al*. (2004, *apud* ALMEIDA, 2012), os testes de dependência espacial podem ser divididos entre os testes difusos e os testes focados. Os testes difusos são aqueles que, ao mesmo tempo em que rejeitam a hipótese nula de que os resíduos espaciais são independentes contra a hipótese alternativa de dependência espacial, não indicam qual a autocorrelação espacial poderia estaria presente. Já os modelos focados fornecem alguma indicação do tipo predominante de autocorrelação espacial, tendo em vista que a mesma depende de diversos fatores, conforme será visto nos itens subsequentes.

#### 3.5.2.1 Testes difusos de dependência espacial

O principal teste difuso de dependência espacial é o índice de Moran, o qual mostra um valor de autocorrelação espacial sem indicar o modelo espacial que poderia ser aplicado como forma de detectar tal dependência. Além dessa limitação, o índice de Moran pode estar indicando a dependência espacial quando na realidade não está ocorrendo. Nesse caso, pode ser devido à não normalidade dos erros, pela presença de heterocedasticidade ou má especificação do modelo, sendo também muito dependente da matriz de ponderação espacial (ALMEIDA, 2012).

Outro teste difuso apresentado por ALMEIDA (2012) é o teste proposto por

KELEJIAN e ROBINSON (1992) denominado de teste Kelejian-Robinson (KR), o qual apresenta a vantagem em relação ao índice de Moran de não pressupor a normalidade dos resíduos da regressão, a de poder ser aplicado a modelos lineares e não lineares e a de não demandar matriz de ponderação espacial. Como desvantagem, está a de ser mais apropriado para grandes amostras e a de ser sensível somente aos vizinhos de primeira ordem. Mais detalhes sobre o teste podem ser vistos em ALMEIDA (2012).

#### 3.5.2.2 Testes focados de dependência espacial

Os testes focados, conforme mencionado anteriormente, são aqueles que fornecem uma indicação da autocorrelação espacial presente nos erros. No entanto, não são capazes de indicar o mais adequado dentre todos os modelos espaciais possíveis. Os testes de multiplicadores de Lagrange (ML) e sua variação robusta costumam ser empregados para que se possa verificar a existência de autocorrelação espacial na variável dependente (modelos SAR) ou nos erros (modelos SEM).Os modelos espaciais serão apresentados de forma mais detalhada nos itens seguintes. A explicação sobre os testes pode ser obtida em ALMEIDA (2012). Aqui serão apresentados, por questão de simplificação, somente a forma como são aplicados e interpretados. Os testes de multiplicadores de Lagrange (teste ML) podem ser divididos em quatro tipos: ML do modelo SAR, ML robusto do modelo SAR, ML do modelo CAR e ML robusto do modelo CAR. Os testes costumam ser aplicados segundo o seguinte procedimento ilustrado no fluxograma da Figura 13, denominado de procedimento híbrido de especificação de modelos espaciais proposto por ANSELIN (1996).

Como se pode ver no fluxograma, quando não se tem nenhum ML significativo, fica-se com a regressão múltipla. Quando se tem somente um significativo, fica-se com este modelo significativo, independentemente da versão robusta. Quando se tem ML significativo para ambos os modelos, verifica-se a versão robusta e aquele que apresentar a maior significância será o modelo escolhido.

Este procedimento é criticado por ser empregado somente para a diferenciação dos modelos SAR e CAR e poder conduzir a modelos que não os mais representativos. Por outro lado, ANSELIN *et al.* (2004, *apud* ALMEIDA, 2012), mencionam a crítica de Henry quanto a este modelo, pelo fato dos níveis de significância dos testes implementados não serem efetivamente conhecidos e pelo teste ser condicionado por hipóteses arbitrárias e se estas hipóteses foram rejeitadas numa sequência, as inferências

tiradas anteriormente ficam inválidas. No entanto, tendo em vista que na pesquisa se emprega os modelos SAR e CAR, este procedimento será empregado na pesquisa como forma de auxílio na busca do melhor modelo, mas não de forma conclusiva.

ALMEIDA (2012) sugere como critério mais geral a de se aplicar os diversos modelos espaciais e verificar o modelo que melhor se ajusta aos dados dentro de uma coerência com a parte conceitual sobre o assunto.

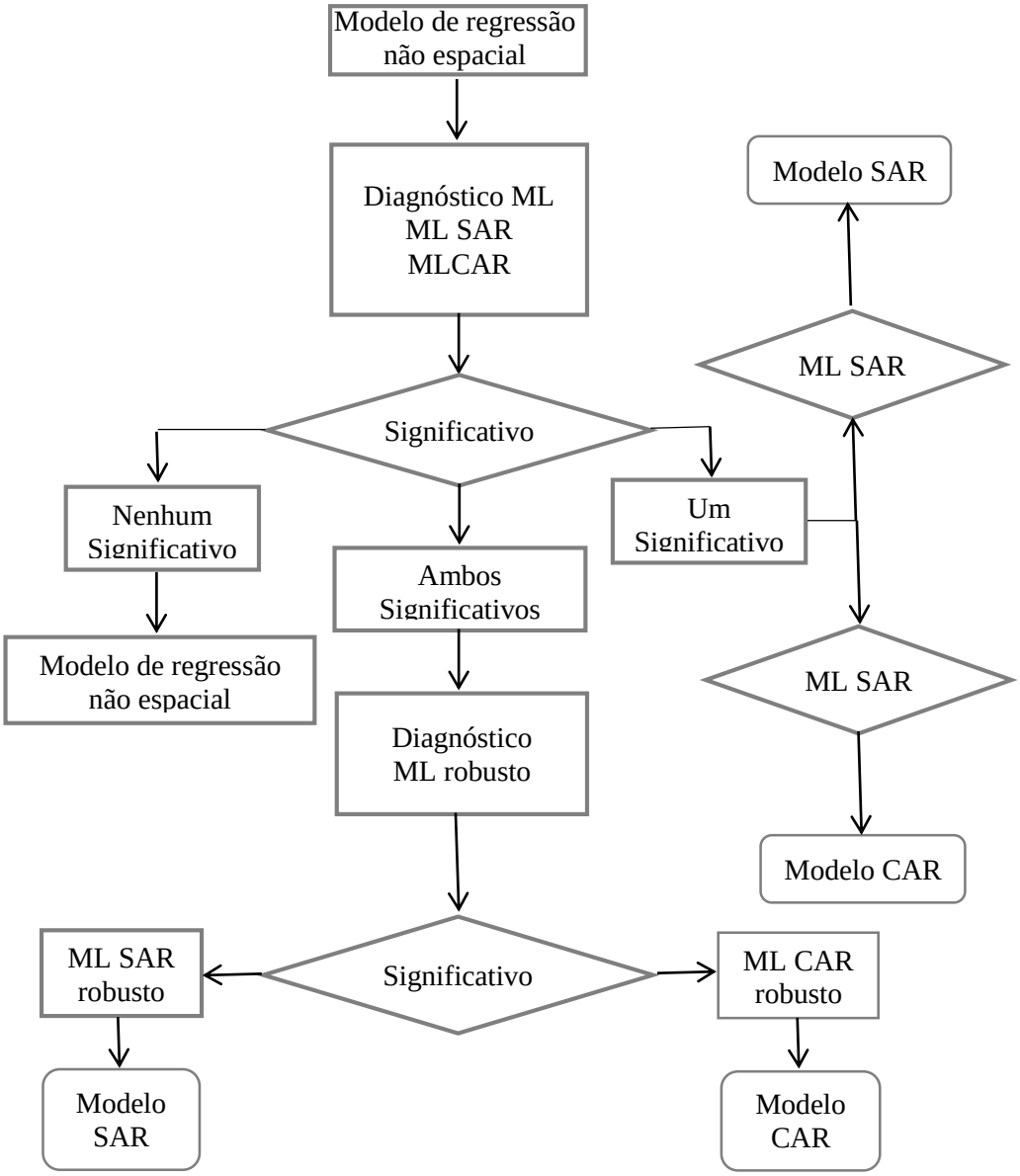


Figura 13 Fluxograma com o procedimento híbrido de especificação de modelos espaciais.
Adaptado de ALMEIDA (2012)

### 3.5.3 Modelos espaciais

Os modelos espaciais serão divididos entre aqueles que consideram a dependência espacial e aqueles que contemplam a heterogeneidade espacial, nesta sequência, pois a heterogeneidade espacial deve ser verificada somente após se ter verificado a existência da dependência espacial.

#### 3.5.3.1 Modelos que contemplam a dependência espacial

Conforme mencionado anteriormente, antes da aplicação dos modelos espaciais, é preciso verificar a dependência espacial por meio dos testes difusos ou focados, os quais poderão dar uma indicação dos tipos de modelos a serem empregados. Como forma de encontrar modelos que possam eliminar ou ao menos atenuar este efeito, deve-se buscar aqueles que, em um determinado nível de agregação, ou seja, desconsiderando por ora a existência os erros de medida, contemplem a questão da interação espacial e da má especificação do modelo. Com essa finalidade, serão aplicados na pesquisa alguns modelos econométricos espaciais, os quais adicionam ao modelo de regressão múltipla um ou mais termos, denominados de defasagens espaciais.

Partindo-se dos modelos de regressão múltipla é possível verificar que as variáveis a serem determinadas são os coeficientes e que as variáveis conhecidas ou obtidas após a construção do modelo são a variável dependente $y$, as variáveis explicativas $X$ e o erro $\varepsilon$. Considerando $W$ como a matriz de vizinhança, a defasagem espacial poderia então estar presente tanto na variável dependente $y$ representada por $Wy$, em uma ou mais variáveis explicativas $X$ representadas por $WX$ ou no erro $\varepsilon$ representado por $W\varepsilon$, podendo estar presente em mais de uma delas ao mesmo tempo, gerando diversos modelos diferentes.

Além disso, estes modelos ainda podem ser modelos globais, quando a dependência espacial está presente em toda a região de trabalho da mesma forma, locais quando se encontra somente em algumas regiões e locais e globais ao mesmo tempo quando apresenta umas variáveis que trazem o efeito global e outras o efeito local.

A Tabela 5 apresenta diversos modelos espaciais apresentados em ALMEIDA (2012) indicando o local onde ocorre a dependência espacial, o alcance dos modelos, estes divididos entre globais (G), locais (L) ou globais e locais (GL).

Tabela 5 Modelos espaciais com o local, alcance dos modelos e transbordamento

| **Modelos espaciais** | **Local da dependência espacial** | **Alcance** |
|---|---|---|
| SAR (Spatial Auto Regressive) | Valores da variável dependente na vizinhança *j* influenciam a variável dependente *i* | G |
| CAR (Conditional Autoregressive) | Valores do erro na vizinhança *j* influenciam o erro em *i* | G |
| SAC | Valores da variável dependente e do erro na vizinhança *j* influenciam a variável dependente e o erro em *i* | G |
| SMA (Spatial Moving Average) | Valores do erro na vizinhança *j* influenciam o erro em *i*, em alguns locais da área de estudo (L). | L |
| SLX | Valores de uma ou mais variáveis explicativas na vizinhança *j* influenciam a variável dependente em *i*. | L |
| SLXMA | Valores de uma ou mais variáveis explicativas e no erro na vizinhança *j* influenciam a variável dependente e o erro em *i*. | L |
| SDM (Spatial Durbin Model) | Valores de uma ou mais variáveis explicativas (L) e da variável dependente (G) na vizinhança *j* influenciam as variáveis explicativas e dependente em *i*. | GL |
| SARMA | Valores da variável dependente (G) e do erro na vizinhança *j*(L) influenciam a variável dependente e o erro em *i* | GL |
| SDEM | Valores de uma ou mais variáveis explicativas (L) e no erro na vizinhança *j* (G) influenciam a variável dependente e o erro em *i*. | GL |

Fonte: Adaptado de ALMEIDA (2012)

Estes modelos podem ser aplicados a outros modelos que não somente os modelos de regressão múltipla, como é o caso dos modelos CAR aplicados em modelos lineares generalizados ou modelos com abordagem bayesiana.

ALMEIDA (2012) ressalta que, embora a inclusão das defasagens possa estar correta, pode haver ainda inconsistência na estimação dos coeficientes caso o método de estimação do modelo não esteja correto. Na pesquisa, empregou-se o método dos mínimos quadrados ordinários para a regressão múltipla e de máxima verossimilhança para os demais modelos.

Por questões didáticas, serão apresentados aqui somente os modelos mais simples, nos quais a dependência espacial está contemplada pela variável dependente (SAR), erro

global (SEM), erro local (SMA) e pela variável explicativa (SLX). Os demais modelos podem ser vistos em ANSELIN (1988) e ALMEIDA (2012).

#### 3.5.3.1.1Alguns modelos de dependência espacial

O modelo espacial SAR, do inglês *Spatial Autoregressive* ou *Spatial Lag* é um modelo no qual se acrescenta um termo ao lado direito da equação em função do valor da variável dependente na vizinhança *WY*, sendo representado pela Eq. 13.

$$Y = \rho WY + \beta X + \varepsilon \qquad \text{Eq. 13}$$

Onde *W* é a matriz de proximidade espacial e ρ é o coeficiente espacial autoregressivo ou parâmetro de regressão espacial, associado aos valores da autocorrelação espacial da variável $Y_i$ na área $A_i$, com os de $Y_j$ na área $A_j$ contida na vizinhança de *i*. Quando ρ = 0, não existe dependência espacial. $\varepsilon$ representa um erro aleatório com média 0 e variância constante.

O modelo SEM, do inglês *Spatial Error Model*, ou CAR, do inglês *Conditional Autoregressive*, é um modelo em que a dependência espacial está presente no resíduo da regressão em função de não ter sido contemplado pelo modelo. Esse erro com dependência espacial, por sua vez, não pode ser correlacionado com nenhuma variável explicativa do modelo de regressão. Este modelo está representado pela Eq. 14.

$$Y = \beta X + \xi \qquad \text{Eq. 14}$$

Sendo $\xi$ representado pela Eq.15:

$$\xi = \lambda W\xi + \varepsilon \qquad \text{Eq. 15}$$

Onde $\lambda$ representa o coeficiente espacial autoregressivo, $\lambda W\xi$ é a o vetor de erro ponderado espacialmente e $\varepsilon$ é a componente do erro com variância constante e não correlacionada. Quando $\lambda = 0$ não há correlação espacial. Cabe reforçar que o erro possui alcance global assim como no modelo SAR, ou seja, o impacto da multiplicação da variável pela matriz de ponderação está presente em toda a área de estudo.

O modelo SMA é um modelo que apresenta uma formulação muito próxima ao modelo SEM, sendo que o erro espacial agora segue um processo de média móvel de primeira ordem. É representada pela Eq. 16.

$$Y = \beta X + \xi \qquad \text{Eq. 16}$$

Sendo $\xi$ representado pela Eq.17.

$$\xi = \gamma W\varepsilon + \varepsilon \qquad \text{Eq. 17}$$

Onde $\gamma$ representa o coeficiente de média móvel espacial, estando também entre os intervalos -1 e 1.

O modelo SLX denominado de modelo regressivo cruzado espacial acrescenta um termo ao modelo referente à influência na variável explicativa presente em *j* em relação a um local *i*. É representado pela variável pela Eq. 18.

$$Y = \beta X + WX\tau + \varepsilon \qquad \text{Eq. 18}$$

Onde $\tau$ representa um vetor e não um escalar de forma que nem todas as variáveis explicativas precisam ser incluídas no modelo.

#### 3.5.3.2 Modelos que contemplam a heterogeneidade espacial

Conforme mencionado anteriormente, a heterogeneidade espacial está presente em uma região onde se observa que o comportamento do fenômeno em estudo não é homogêneo em toda ela. Tal ideia parece razoável em locais onde existem grandes desigualdades sociais e econômicas, como é o caso do município de Rio de Janeiro. A heterogeneidade pode ser observável ou não observável. A heterogeneidade utilizada na pesquisa será a heterogeneidade observável. A não observável está presente em modelos em que a heterogeneidade está presente, mas não é contemplada por nenhuma variável presente no modelo, sendo atribuída a uma variável que não pode ser observada. É importante não confundir, portanto, variável não observada com variável observável, mas omitida do modelo.

A heterogeneidade espacial observável pode ser verificada sob três aspectos: heterogeneidade no intercepto, nos coeficientes de inclinação do modelo de regressão e no erro.

##### 3.5.3.2.1 Heterogeneidade no intercepto

Na verificação da heterogeneidade do intercepto busca-se encontrar a existência de algum tipo de comportamento nos dados que faça com que a ordem de grandeza de um conjunto de dados referentes ao fenômeno em estudo possa ser diferente de outro conjunto de dados. No caso do intercepto, essa diferença seria causada não pelas variáveis

explicativas, mas sim pelo intercepto. Na prática, aplicam-se técnicas para detectar alguma grandeza que represente essa "defasagem" que estaria presente no conjunto de dados. Como exemplos dessas técnicas estariam as técnicas de ANOVA Espacial e modelos de superfície de tendência. Estas técnicas serão empregadas na pesquisa na análise exploratória como forma de verificar a existência de regimes espaciais.

No caso da ANOVA espacial, essa "defasagem" estaria presente na diferença entre a média da variável em estudo e o valor da variável em cada aglomerado, sendo obtida por mínimos quadrados ordinários a partir de um modelo de regressão simples representada pela Eq. 19.

$$Y = \beta_1 + \beta_2 X_1 + \varepsilon \qquad \text{Eq. 19}$$

Onde a variável explicativa $X_1$ é uma variável *dummy*, ou seja, possui somente os valores 0 ou 1 e $\varepsilon$ é um erro aleatório com média 0 e variância constante. Uma variável *dummy* significativa indica a existência de um diferença não desprezível entre o comportamento da variável independente $\beta_1$em um dado local e na média da região.

No caso dos modelos de superfície de tendência, essa "defasagem" estaria presente nos valores da posição espacial dos dados, sendo traduzido comumente pelos valores latitude e longitude. Este modelo é aplicado na região como um todo, sendo representado pela Eq. 20.

$$Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \varepsilon \qquad \text{Eq. 20}$$

Onde $\beta_2$ seria o valor da longitude e $\beta_3$ o valor da latitude. Caso os valores destes coeficientes sejam positivos e significativos, existe uma tendência de aumento da variável com a longitude ou latitude, respectivamente, ocorrendo o inverso no caso dos sinais serem negativos. Estes modelos também podem ser representados na sua forma quadrática.

Segundo ALMEIDA (2012), os modelos de superfícies de tendência costumam contemplar grande parte da variação da autocorrelação espacial, podendo até mesmo eliminar em grande parte a autocorrelação espacial dos resíduos. Na medida em que capta a parte determinística da tendência espacial dos dados, não contempla a componente estocástica, de menor valor e mais localizada.

Embora seja um interpolador determinístico com efeito espacial contínuo e aqui se esteja trabalhando com dados discretos, pode ser empregado com o objetivo de fazer uma análise exploratória do comportamento da média geral da regressão $\beta_1$, não se querendo

reduzir o modelo ao emprego exclusivo de variáveis associadas à localização, podendo ser empregado para reduzir ou eliminar o efeito da dependência espacial. Esse modelo é importante na medida em que pode suavizar os dados por meio da retirada da "defasagem" atribuída à superfície de tendência.

#### 3.5.3.2 Heterogeneidade nos coeficientes e no intercepto

No caso da heterogeneidade no intercepto, essas "defasagens" entre regiões seriam atribuídas aos coeficientes, o que faz com que não tenham a mesma ordem de grandeza, assim como foi verificado para os interceptos. Tal variação pode ocorrer de forma discreta ou contínua. No caso da discreta, parte-se de um conjunto de dados que esteja separado de outro conjunto de dados por algum critério que pode ser determinado por análise exploratória. Nesse caso, os coeficientes apresentam um valor fixo em cada um dos regimes espaciais. Como exemplo, tem-se a separação entre os considerados ricos ou pobres quanto a algum indicador associado ao fenômeno em estudo. Nesse caso ocorrem mudanças tanto no intercepto como nos coeficientes do modelo.

A heterogeneidade tanto no intercepto como nos coeficientes nesta pesquisa empregará o modelo de regimes espaciais. Neste modelo, os coeficientes e o intercepto apresentam valores diferentes para cada regime espacial. Cada regime espacial representa uma sub-região da área total de trabalho. O somatório de todos os n regimes espaciais equivale à área de toda a região.

A representação matricial de *n* regimes espaciais em uma dada região com *m* variáveis explicativas é dada pela Eq. 21.

$$\begin{bmatrix} y_1 \\ \dots \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} X_{11} & 0 & \dots & X_{mn} \\ 0 & \dots & 0 & \dots \\ \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & X_n \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \dots \\ \dots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \dots \\ \dots \\ \varepsilon_n \end{bmatrix} \qquad \text{Eq. 21}$$

Onde se pode observar que cada regime espacial *n* apresenta um valor diferente da variável dependente, variável explicativa, coeficiente e erro. Considerando o erro como tendo uma distribuição normal com média 0 e variância Ω, a variância é um valor constante no caso dos regimes espaciais homocedásticos e diferente para cada regime espacial que contemple a heterocedasticidade.

Um teste a ser aplicado na pesquisa para verificar a existência de estabilidade nos coeficientes e intercepto é o teste Chow espacial. Este teste compara a soma dos

quadrados dos resíduos da regressão de todo o conjunto de dados com a soma dos quadrados dos resíduos dos subconjuntos de dados no qual foi dividido todo o conjunto de dados. Mais detalhes podem ser vistos em ALMEIDA (2012).

Segundo ANSELIN (1990, *apud* ALMEIDA, 2012), a autocorrelação espacial pode invalidar o teste de Chow espacial, devendo ser incorporada no modelo por meio de um dos modelos previstos e que contemplam a dependência espacial.

Quando essa heterogeneidade nos coeficientes se manifesta de forma diferente no espaço, podem-se aplicar outros modelos, como é o caso dos modelos de expansão espacial proposto por CASETTI (1972) e regressão ponderada espacialmente, os quais não serão abordados neste trabalho. Mais detalhes sobre estes modelos podem ser vistos em ALMEIDA (2012). Nestes modelos, ao contrário dos regimes espaciais cujos coeficientes são constantes em alguns conjuntos de dados, os coeficientes variam em todo o conjunto de dados.

#### 3.5.3.3 Heterogeneidade no erro ou heterocedasticidade

Os modelos espaciais serão empregados após serem aplicados alguns testes e serem verificados nos resíduos da regressão a dependência e a heterocedasticidade. Os testes de dependência espacial foram vistos anteriormente. Caso haja dependência espacial, a mesma deve ser corrigida primeiramente por influenciar sobremaneira a heterocedasticidade. Após modelar a dependência espacial, deve-se então verificar se a heterocedasticidade foi corrigida. Caso persista, ambos os efeitos espaciais devem ser modelados conjuntamente.

ANSELIN (2005) indica alguns testes para detectar a heterocedasticidade. São eles os testes de Breusch-Pagan e Koenker-Bassett, os quais assumem uma forma definida para a heterocedasticidade.

Nestes testes, o objetivo é inferir a variância em cada local *i* em função das variáveis explicativas, ou seja, $\sigma^2_i = E(u^2_i \,|x_i)$, onde $u_i$ são os erros da regressão, os quais não são observados mas estimados por mínimos quadrados ordinários. Logo, o valor estimado do erro em cada local $\hat{u}_i$ é representado pela Eq. 22.

$$\hat{u}_i = y_i - \hat{y}_i \qquad \text{Eq. 22}$$

Considerando o $\hat{u}_i$ como sendo um estimador centrado para $\sigma^2_i$ quando $\sigma^2_i \neq \sigma^2$, então $\hat{u}_i$seria um valor *proxy* de $\sigma^2_i = E(u^2_i \,|x_i)$. Dessa forma, a hipótese nula dos erros

homocedásticos seria dada pela Eq. 23.

$$H_0 = \sigma_i^2 = E(u_i^2/x_i) = E(u_i^2) = \sigma^2 \quad \text{Eq. 23}$$

A estatística de teste é feita baseando-se em um modelo de regressão múltipla onde $\sigma^2_i = u_i^2 = f(x_i)$. No caso do teste de Breusch-Pagan, a equação ficaria conforme a Eq. 24.

$$u_i^2 = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \cdots + \beta_k X_{k-1} + v_i \quad \text{Eq. 24}$$

Onde a hipótese nula $H_0$ seria a de $\beta_1 = \ldots = \beta_k$ e a hipótese alternativa seria que pelo menos um dos coeficientes fosse diferente de zero. Embora comumente se empregue a forma linear, o programa GeoDa, empregado nesta pesquisa, adota uma função do quadrado das variáveis explicativas, o que leva a resultados um pouco diferentes. O teste de Koenker-Bassett seria um teste de Breusch-Pagan com os resíduos studentizados. O teste de White (1980), por sua vez, também chamado de teste de especificação robusta de heterocedastidade apresenta outras formas que não a linear, tais como quadrados e produtos cruzados. Como exemplo para $k$=3, tem-se a Eq. 25.

$$u_i^2 = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_1^2 + \beta_5 X_2^2 + \beta_k X_1 X_2 + v_i \quad \text{Eq. 25}$$

Onde $v_i$ é o resíduo da regressão em um local $i$.

Uma vez detectada a heterogeneidade espacial, poderão ser aplicados diversos modelos que contemplem a heterogeneidade no erro, os quais não serão aqui abordados tendo em vista que no contexto da tese serão aplicados somente os modelos com heterogeneidade dos coeficientes e no intercepto. Tais modelos poderão ser encontrados nos livros de Econometria Espacial.

### 3.5.4 Verificação da qualidade do ajuste

De forma a verificar o quão ajustados aos dados reais estão os resultados estimados pelo modelo, pode-se empregar diversas medidas, como é o caso do desvio absoluto médio – MAD - do inglês *Mean Absolute Deviation* (XU *et al.*, 2014) representado pela Eq. 26.

$$MAD = \frac{\sum_1^n |y_i - \hat{y}_i|}{n} \quad \text{Eq. 26}$$

Onde n é o número de observações $y_i$ é o valor real da variável e $\hat{y}_i$ representa o valor estimado da variável $y$.

Outra forma é pelo erro quadrático médio preditivo - MSPE - do inglês *Mean Squared Predictive Error* (XU *et al.*, 2014), representado pela Eq. 27.

$$MSPE = \frac{\sum_1^n (y_i - \hat{y}_i)^2}{n} \quad \text{Eq. 27}$$

Onde *n* é o número de observações $y_i$ é o valor real da variável e $\hat{y}_i$ representa o valor estimado da variável *y*.

A raiz quadrada do MSPE denomina-se erro médio quadrático – RMS ou RMSE, do inglês *Root Mean Square Error ou RMSD*, do inglês *Root Mean Square Deviation*, também chamado de desvio médio quadrático dado pela Eq. 28.

$$RMS = \sqrt{\frac{\sum_1^n (y_i - \hat{y}_i)^2}{n}} \quad \text{Eq. 28}$$

Onde $y_i$ é o valor real da variável em *i* e $\hat{y}_i$ representa o valor estimado da variável *y*.

Uma forma gráfica de verificação da qualidade do ajuste pode ser feita por meio da aplicação do método dos resíduos acumulados, também chamado de CURE (do inglês CUmulative REsiduals). O CURE consiste em um gráfico em que se tem no eixo das abscissas a variável de interesse e o eixo das ordenadas o resíduo acumulado de cada n, a partir do ordenamento dos N resíduos em ordem crescente da variável de interesse. Considerando que $\hat{\sigma}^2(n)$ seja o somatório dos resíduos de 1 até n e $\hat{\sigma}^2(N)$ o somatório do quadrado dos resíduos de todos os N resíduos obtidos pelo modelo. O valor de $\sigma^*$está mostrado da Eq. 29.

$$\sigma^* = \sqrt{1 - \frac{\hat{\sigma}^2(n)}{\hat{\sigma}^2(N)}} \quad \text{Eq. 29}$$

Considerando ± $2\sigma^*(n)$ são os limites do gráfico CURE para cada n. O bom CURE situa-se em torno do zero e o mau CURE totalmente acima ou abaixo de zero.

Como forma de comparar o resultado obtido por mais de um modelo, costuma-se empregar o critério de informação de Akaike, que pode ser utilizado para determinar a qualidade do ajuste do modelo estatístico adotado sendo expresso pela Eq. 30.

$$\text{AIC} = -2 * \text{LIK} + 2k \quad \text{Eq. 30}$$

Onde LIK é o log de verossimilhança maximizado e *k* é o número de coeficientes de regressão. Segundo DRUCK *et al.* (2004), existem outros critérios de informação. Porém a maior parte dos mesmos são variações do AIC, com mudanças na forma de penalização de parâmetros ou observações. Quanto à interpretação do resultado, quanto menor o valor do AIC, melhor o modelo.

### 3.5.5 Validação do modelo

A validação envolve o processamento do modelo com os parâmetros determinados durante a fase de calibração e com dados não utilizados na calibração e comparar as previsões obtidas nessa etapa com as obtidas na calibração.

A validação dos dados de acidentes nesta pesquisa será feita construindo-se um intervalo de previsões de novas observações da densidade de acidentes. Para o caso da regressão múltipla, este intervalo é dado pela Eq. 31. (MONTGOMERY e RUNGER, 2012):

$$Y_i \leq \hat{y}_i + t_{\frac{\alpha}{2},n-p}\sqrt{\hat{\sigma}^2(1 + x_i^T(X^TX)^{-1}x_i)}$$
$$\hat{y}_i - t_{\frac{\alpha}{2},n-p}\sqrt{\hat{\sigma}^2(1 + x_i^T(X^TX)^{-1}x_i)} \leq Y_i \qquad \text{Eq. 31}$$

Onde *t* é uma distribuição de *Student* com *n-p* graus de liberdade, n é o número de observações de cada variável e *p* o número de variáveis explicativas *k* somado com 1.

MIRANDA-MORENO *et al.* (2011) empregaram a correlação de Pearson para comparar os resultados dos modelos obtidos na validação com os valores observados da variável resposta. Empregam também a medida do erro quadrático médio normalizado, do inglês *Normalized Root Mean Square Deviation*, o qual é obtido dividindo o RMS pelo *range* dos valores da variável observada, representado pela Eq. 32.

$$NRMS = \frac{\sqrt{\frac{\sum_1^n (y_i - \hat{y}_i)^2}{n}}}{y_{max} - y_{min}} \qquad \text{Eq. 32}$$

## 4. METODOLOGIA PROPOSTA

O presente capítulo apresenta a metodologia proposta para a análise espacial de acidentes de trânsito, a qual contempla três grandes etapas: aquisição (AQ), compreensão da distribuição espacial (DE) e modelagem dos dados de acidentes (MO). A etapa de aquisição compreende a caracterização da área de estudo, a qual serve de suporte para a definição de variáveis a serem obtidas e as análises a serem feitas, a coleta e preparação dos dados, bem como a verificação dos efeitos de bordas. Na etapa de compreensão da distribuição espacial estão a análise visual e a análise exploratória dos acidentes. Por fim, na etapa de modelagem estão a seleção das variáveis explicativas, a calibração dos dados estatísticos, a análise da heterogeneidade espacial, a geração e análise dos dados para verificar a questão do MAUP e a validação dos modelos. O fluxograma da Figura 14 contém as etapas constantes na metodologia proposta.

| | | |
|---|---|---|
| AQ | Caracterização da área de estudo | 4.1 |
| | Preparação dos dados | 4.2 |
| DE | Verificação dos efeitos de bordas | 4.3 |
| | Análise visual | 4.4 |
| | Análise exploratória | 4.5 |
| MO | Seleção das variáveis explicativas | 4.6 |
| | Calibração dos dados estatísticos | 4.7 |
| | Análise da heterogeneidade espacial | 4.8 |
| | MAUP –geração de dados | 4.9 |
| | MAUP – análise de sensibilidade | 4.10 |
| | Validação dos modelos | 4.11 |

Figura 14 Fluxograma com as etapas da metodologia proposta

## 4.1 Caracterização da área de estudo

A caracterização da área de estudo é uma etapa de prospecção de conhecimento e que serve de subsídio para a escolha das variáveis que poderão ser empregadas na pesquisa. Pode também fornecer valioso auxílio às etapas de análise exploratória e na seleção de modelos, fornecendo informações que possam ajudar na interpretação dos resultados das análises espaciais.

## 4.2 Preparação dos dados

Nesta fase, são obtidos os dados de acidentes e as variáveis explicativas e também são preparados para serem utilizados nos diversos programas empregados na metodologia.

Ao se manipular os dados geográficos, é importante estar atento para dois fatores: o sistema geodésico e a escala dos dados. Quanto ao sistema geodésico, é fundamental conhecer o sistema em que cada um dos dados foi obtido para que, caso seja solicitado, possa ser fornecido ao aplicativo de SIG. Essa informação, juntamente com os parâmetros de conversão entre os sistemas geodésicos são importantes para que as feições possam ser exibidas no local correto. Estes parâmetros são fornecidos pelo IBGE e podem ser facilmente encontrados na sua página eletrônica. Caso tais parâmetros não tenham sido inseridos ou tenham sido feitos de forma incorreta, será perceptível certa "defasagem" sistemática das feições deste conjunto de feições em relação à posição correta.

Quanto à escala dos dados, quando se trabalha com dados em escalas consideravelmente diferentes e se superpõe os mesmos em uma mesma região geográfica, observa-se que os mesmos não coincidem e neste caso não ocorrerá somente uma defasagem entre as feições mas uma diferença não sistemática entre os traçados das feições. Para evitar tal problema, é importante que os dados estejam em escalas senão iguais ou, ao menos, próximas. No entanto, conseguir que os dados estejam em escalas próximas muitas vezes não é simples, principalmente em regiões em que haja maior carência de informações geográficas. Para tentar contornar tal situação, deve-se antes de mais nada definir a escala de trabalho ou ao menos o nível de agregação em que os dados serão trabalhados no início da pesquisa e buscar que esse nível de agregação possa ter uma precisão compatível com o nível de agregação menos detalhado dentre as variáveis empregadas na análise. Como exemplo, pode-se citar as variáveis quantidade de

habitantes em um dado local, a qual se encontra no nível de setor censitário, e o número de empregos, que está no nível geográfico de bairros. Para que se possa utilizar as mesmas variáveis em um mesmo modelo, deve-se utilizar o nível geográfico dos bairros pois é o nível mais desagregado que contempla ambos os níveis de informação. Neste caso, o fato dos setores censitários estarem associados aos bairros faz com que os limites dos bairros possam ser obtidos pela agregação dos setores censitários.

No entanto, no caso em que se tem variáveis obtidas nos níveis de agregação cujos limites não coincidem entre si e/ou com o nível de agregação adotado na análise, é preciso criar um critério que possa ser adotado para associar os valores da variável para o nível de agregação adotado na pesquisa. Um critério comumente empregado é o de ponderar o valor da variável em uma dada área pelo valor da proporção desta área que superpõe o nível de agregação em que se deseja obter o valor dessa variável.

Para o caso dos dados tabulares, é importante verificar o nível geográfico em que foram coletados para que não se corra o risco de associar os dados em um nível de agregação diferente daquele no qual foram obtidos.

Ao se associar os dados pontuais às áreas, é importante saber sobre o nível de precisão com que foram adquiridos. Um erro comum é o de empregar coordenadas obtidas a partir de rastreadores GPS sem considerar a imprecisão presente nestes dados. Tal problema pode se tornar ainda mais crítico quando se trabalha com áreas pequenas onde aumenta a chance dos dados estarem situados nas bordas da área, conforme será detalhado posteriormente.

Em todos os casos, é importante estar atento para a verificação da fonte dos dados e do período de tempo em que foram obtidos. Neste sentido, é indicado que os dados sejam da mesma época ou, ao menos, de períodos de tempo próximos. Caso haja um período de tempo considerável entre as datas de coleta das variáveis envolvidas na modelagem, talvez seja o caso de se aplicar um percentual de correção dos valores a partir das datas mais defasadas baseando-se na variação da série histórica destes dados.

Para o caso dos dados de acidentes, estes poderão ser fornecidos de diversas formas, desde a descrição do local do acidente até em coordenadas obtidas por rastreadores GPS. No caso das informações obtidas pelo endereço, podem-se aplicar as técnicas de geocodificação no sentido de converter em coordenadas uma descrição alfanumérica do local do acidente.

É importante destacar que nessa etapa estão envolvidas numerosas operações espaciais até que se obtenham os valores das variáveis envolvidas no nível de agregação

utilizado na análise. Muitas vezes, a obtenção de uma variável envolve uma longa sequência de operações, onde o resultado de cada uma das operações torna-se o insumo para a operação subsequente. Logo, um erro em uma operação pode invalidar toda a sequência de operações. Nesse momento, é preciso ser muito criterioso com os resultados obtidos após cada uma das etapas. Um dos erros mais comuns é trabalhar com unidades de medidas erradas. Por exemplo, quando se solicita ao programa ArcGIS o cálculo da área de um dado polígono de bairros que se encontra em coordenadas geográficas (latitude e longitude) o mesmo fornece um resultado em uma unidade inexistente que seria a de graus ao quadrado. Por isso, antes de começar a se aplicar as operações espaciais e quando se trabalha em uma região cuja extensão é pequena em relação à largura de um fuso UTM (cuja largura na linha do Equador está em torno de 600 km), como é caso de uma cidade, deve-se converter todos os dados que possam estar em coordenadas geográficas para coordenadas métricas. O principal sistema de coordenadas adotado no Brasil é o sistema *Universal Transversa de Mercator* (UTM). Desse modo, todas as medidas de distância e de área estarão em unidades métricas como quilômetro e quilômetro ao quadrado, respectivamente.

Outro erro que pode ocorrer é a não atualização dos valores das medidas das feições após a aplicação de uma operação espacial. Como exemplo, pode-se citar a extensão das vias de uma cidade após serem cortadas pelos limites dos polígonos das áreas dos bairros, as quais podem continuar assumindo os valores antes da operação de corte.

Após esta etapa, todas as variáveis devem estar associadas ao nível de agregação adotado na pesquisa, para que sejam devidamente analisadas e, se for o caso, empregadas na modelagem. Cabe ressaltar que as variáveis, tal como foram obtidas, também podem ser úteis na análise visual dos dados geográficos, podendo ajudar na elucidação do fenômeno em estudo.

### 4.3 Verificação dos efeitos de bordas dos acidentes

Conforme mencionado anteriormente, ao se obter dados pontuais, como é o caso dos acidentes de trânsito obtidos por meio de rastreadores GPS ou por qualquer tipo de geocodificação, é preciso atentar para a imprecisão destes dados, o que pode fazer com que sejam associados a feições a que na realidade não pertençam, simplesmente por uma diferença entre as escalas dos dados GPS e dessas feições. Para o caso em que se estas feições sejam áreas, esses efeitos são mais sentidos quando se tem polígonos menores,

onde a chance dos pontos ficarem próximos às bordas aumenta em relação aos polígonos maiores. Por exemplo, não há maiores problemas em se associar uma coordenada obtida por GPS ao mapa de um município por este cobrir uma área geográfica muito superior à imprecisão do sistema GPS. No entanto, ao se associar tais dados a um setor censitário de uma área urbana, o qual se encontra em uma escala próxima a uma escala cadastral, é preciso verificar a imprecisão da localização do ponto para que não haja o risco de que seja associado a uma área vizinha da área real. Caso o número destes pontos seja representativo em relação ao número total de pontos, é preciso criar um critério para levar em consideração tal incerteza.

Como forma de verificar tal incerteza, pode-se construir *buffer zones* em torno dos limites das áreas, cuja largura será função da precisão do sistema GPS, e selecionar o número de pontos situados dentro destas zonas. Por fim, compara-se o número de acidentes presente nesta região das *buffer zones* com o total de acidentes de cada área. O percentual de pontos nas bordas pode ser tão expressivo em relação ao número total de pontos de cada região que talvez seja necessário dar um tratamento separado para ambas as regiões. Tal fato foi verificado por SIDDIQUI e ABDEL-ATY (2012), os quais construíram um modelo híbrido, considerando tanto a região interior como a das bordas das áreas.

### 4.4 Análise visual

Na análise visual, busca-se conhecer o espaço geográfico no qual ocorrem os acidentes utilizando-se da visualização espacial da região em estudo. Nesse sentido, auxilia na compreensão da distribuição espacial das variáveis envolvidas na modelagem por meio da contextualização das mesmas na região de estudo, podendo ajudar até mesmo na identificação de novas variáveis explicativas.

Nesta etapa, diversas informações levantadas na caracterização da área de estudo e que não estão envolvidas diretamente na modelagem podem ser empregadas para uma melhor compreensão do fenômeno em estudo. Como exemplo, para o caso da cidade do Rio de Janeiro, pode ser importante visualizar espacialmente o relevo e os polígonos das áreas verdes pelo fato desta ser uma cidade caracterizada por grandes áreas verdes e montanhosas.

No que diz respeito às técnicas de visualização, aqui podem ser aplicadas quaisquer técnicas de visualização espacial de modo a encontrar algum tipo de padrão nos dados

geográficos empregados na pesquisa, podendo ser estáticas ou dinâmicas. O tipo de análise também depende muito se os mesmos são pontuais, lineares ou em área. Por exemplo, para o caso dos dados pontuais, poderão ser geradas superfícies que poderão ajudar a fornecer uma visão diferente daquela obtida a partir da agregação destes dados pontuais em área. Para o caso dos dados de área, costuma-se utilizar os mapas temáticos, com destaque para os coropléticos. Tendo em vista que as variáveis são grandezas quantitativas, podem-se utilizar diversos métodos para a visualização dos mesmos, tais como os métodos dos quartis, do desvio-padrão e da quebra natural. Estes mapas, quando analisados conjuntamente com estatísticas e gráficos, serão de grande valia na análise exploratória das variáveis envolvidas na modelagem.

Esta etapa se confunde com a de análise exploratória, por explorar da mesma forma os dados em busca de identificar padrões nos mesmos. Também pode causar confusão o fato de se obter diversos mapas na análise exploratória tais como o mapa de Moran e *box map*, os quais envolvem da mesma forma a análise visual dos dados. No entanto, no contexto desta metodologia, a análise visual pode ser considerada como uma etapa mais prospectiva e não uma forma de visualizar as diferentes estatísticas obtidas nesta etapa. Será priorizada nesta fase a análise visual da variável dependente e a prospecção de novas variáveis explicativas, além daquelas que porventura já tenham sido obtidas na coleta e preparação dos dados.

Conforme mencionado, após a análise visual, diversas técnicas estatísticas que se utilizam da posição dos dados geográficos serão utilizadas com a finalidade de encontrar padrões e relacionamentos nos mesmos, auxiliando dessa forma a etapa de modelagem.

## 4.5 Análise exploratória

A análise exploratória visa encontrar algum padrão ou relacionamento nos dados por meio de estatísticas, gráficos e mapas, que possam ser úteis para a fase de modelagem.

De forma a verificar a distribuição dos dados e a existência de *outliers* globais, determina-se inicialmente o gráfico do diagrama de caixa (*boxplot*) e respectivo *box map*.

Com o objetivo de determinar a média móvel local e os índices globais e locais de autocorrelação espacial, deve-se antes levar em consideração a ideia de vizinhança, a ser materializada pela matriz de vizinhança.

A escolha da matriz de vizinhança costuma produzir considerável alteração tanto na análise exploratória espacial como nos modelos estatísticos espaciais e por isso deve ser

escolhida com bastante critério. Para tal, é importante verificar a sensibilidade dos resultados em diferentes matrizes de vizinhança. Uma das formas de fazer tal comparação é produzir a cada duas matrizes de vizinhança uma matriz de erro com os resultados dos valores do índice de Moran local (LISA)(ALMEIDA, 2012). Quanto maior for a quantidade de valores Alto-alto, Alto-baixo, Baixo-Alto, Baixo-baixo e Não significativo coincidentes nas matrizes, mais estáveis tendem a ser as matrizes de proximidade.

A partir da escolha da matriz de vizinhança, determina-se o diagrama de espalhamento de Moran no qual se busca observar a existência de *outliers* locais e de pontos de alavancagem.

A análise exploratória espacial reveste-se de grande importância não somente na análise do padrão espacial e da dependência espacial, mas também na verificação da possibilidade da existência da heterogeneidade espacial, por meio da identificação de regimes espaciais. ALMEIDA (2012) sugere a comparação entre um mapa de desvio padrão e um diagrama de dispersão de Moran. Nesse caso, os diferentes quadrantes do diagrama de Moran poderiam identificar possíveis regimes espaciais. Cabe ressaltar que nesta etapa, os testes de dependência espacial são empregados sobre as variáveis. Na etapa de calibração dos modelos estatísticos serão empregadas novamente mas sobre os resíduos dos modelos com a finalidade de verificar a necessidade de se utilizar ou não os modelos espaciais para contemplar a dependência espacial.

Para tornar a divisão da área de estudo em regimes espaciais menos subjetiva, empregou-se nesta pesquisa um método de regionalização, o qual divide a região de estudo no número de clusters que se deseja, baseando-se em um determinado método. Nesta pesquisa será empregado o método de regionalização *REgionalization with Dynamically Constrained Agglomerative Clustering and Partitioning* (REDCAP), o qual será melhor explicado no item da metodologia que versa sobre o problema de MAUP.

Conforme apresentado no capítulo 3, a partir da definição dos polígonos pertencentes a cada regime espacial, pode-se fazer uma ANOVA Espacial com uma variável *dummy*, em que cada valor da variável representa um regime espacial diferente. Por exemplo, no caso de se ter dois regimes espaciais, os valores da variável *dummy* serão 0 para os polígonos de um regime e 1 para o outro regime espacial.

Outra técnica que auxilia no entendimento da distribuição espacial da variável dependente é a aplicação do modelo de superfície de tendência, o qual adota como variáveis explicativas as coordenadas X e Y do centroide dos polígonos.

## 4.6 Seleção das variáveis explicativas

Após terminadas as etapas de análises visual e exploratória, inicia-se uma etapa prévia à modelagem que é a verificação das variáveis explicativas mais correlacionadas entre si e com a variável resposta, utilizando para tal a correlação de Pearson.

Esta fase reveste-se de grande importância, pois é nesta fase que se verifica o comportamento das variáveis sugeridas pela bibliografia ou adotadas a partir da caracterização da área de estudo e da análise visual. Este resultado já fornece por si mesmo uma informação valiosa, que pode ser por ora descartada pela sua alta correlação com outra variável explicativa mais representativa, mas pode vir a ser utilizada em outras modelagens.

Em um primeiro momento, selecionam-se as variáveis que estejam mais correlacionadas com a variável dependente. Em seguida, começando da variável mais correlacionada para a menos correlacionada, observa-se a correlação de cada uma das variáveis com as variáveis explicativas. Considera-se nesta pesquisa como correlacionadas aquelas variáveis cujo coeficiente de correlação de Pearson seja menor que -0,4 ou maior que +0,4. Por fim, eliminam-se as variáveis que sejam correlacionadas entre si ou que possuam baixa correlação com a variável resposta.

Após a seleção das variáveis explicativas pode ser elucidativo plotar um gráfico da variável explicativa em função da variável dependente em uma planilha eletrônica de modo a se ter uma visão da relação entre as variáveis, assim como a verificação da linha de tendência que melhor representa a relação entre as variáveis. Pode-se também gerar mapas temáticos de forma a conhecer melhor a distribuição espacial destas variáveis.

## 4.7 Calibração dos modelos estatísticos

Após a seleção das variáveis que estarão envolvidas na modelagem, inicia-se a fase de calibração, na qual serão testados diferentes modelos, partindo dos modelos mais simples para os mais complexos. Conforme visto na revisão bibliográfica, diversos modelos poderão ser testados, em função das características dos dados com os quais se esteja trabalhando.

Em se tratando dos modelos de regressão múltipla, deve-se verificar se os pressupostos deste tipo de modelo são atendidos (hipóteses de Gauss-Markov). Pode-se ainda testar os modelos lineares generalizados e aqueles com abordagem bayesiana,

dentre outros.

Os valores das variáveis podem ser alterados por algum tipo de transformação, como é o caso da aplicação de uma transformação log, que pode ser aplicada para variáveis cujos valores são diferentes de zero, permitindo que haja uma redução nas variâncias nas variáveis e entre as mesmas, diminuindo o efeito da heterocedasticidade entre as variáveis (GUJARATI, 2003 *apud* QUDDUS, 2008). Este autor menciona ainda como vantagem da utilização da transformação log, a de interpretação dos parâmetros do modelo, que fica sendo igual aos da constante de elasticidade ao contrário de um coeficiente de inclinação.

Nesta pesquisa são considerados ainda os modelos espaciais cujo principal objetivo é contemplar a dependência espacial, conforme visto anteriormente. A dependência espacial pode ser verificada pelos testes difusos nos resíduos do modelo, como é o caso do índice de Moran ou pelos testes focados, empregando os multiplicadores de Lagrange, para o caso dos modelos SAR e CAR. Dependendo do tipo de dependência espacial, diversos modelos espaciais poderão ser testados.

Após eliminar ou ao menos diminuir os valores de dependência espacial para valores próximos de zero, verifica-se a heterogeneidade espacial. Entre os modelos que contemplam a heterogeneidade espacial observável estão aqueles que consideram que a heterogeneidade está presente nos coeficientes e aqueles que consideram como estando presentes no erro.

Por fim, os resultados de todos os modelos deverão ser comparados a partir da análise dos resíduos, de forma a verificar os modelos com melhores ajustes.

Após a seleção do melhor modelo, serão elaborados os modelos considerando agregações diferentes, de modo a verificar a estabilidade da média dos coeficientes das variáveis explicativas e, portanto, contemplar o MAUP.

Nos itens seguintes são apresentados os passos seguidos na análise dos resultados obtidos dos modelos não espaciais, espaciais e posteriormente análise dos regimes espaciais.

Em um primeiro momento a análise do quão ajustado está o modelo em relação aos dados de acidentes é feita observando-se os valores do $R^2$ (modelo de regressão múltipla) ou do critério de informação de Akaike. Os sinais dos coeficientes indicam a relação direta ou inversa entre cada uma das variáveis explicativas e a variável resposta. O nível de significância destas variáveis deve estar abaixo de 0,05, para o caso do nível de confiança desejado ser de 95%. O nível de significância do intercepto, em princípio, deve ser deixado mesmo que esteja abaixo do valor desejado. O mesmo só deve ser eliminado

da modelagem caso se tenha muita certeza da reta de regressão passar pela origem do gráfico. O programa utilizado na modelagem foi o programa GeoDa Space, cuja tela consta na Figura 15. As ferramentas deste programa empregadas na metodologia serão apresentadas como forma de facilitar a compreensão da análise dos resultados. Na parte esquerda constam os dados de entrada e a matriz de ponderação ou pesos, os quais podem ser importados ou criados no programa. Na parte central superior consta a variável resposta (Y) e na parte inferior os regimes espaciais (R). Na parte direita estão as variáveis explicativas. A janela superposta à tela do programa contém a lista de todas as variáveis que constam no arquivo de entrada. Para associar a cada caixa do programa as respectivas variáveis, basta arrastar as mesmas para a caixas. O regime espacial será processado a partir de uma variável gerada a partir do programa REDCAP, conforme será visto mais adiante. Na parte inferior constam os tipos de modelos que poderão ser empregados. Caso se deseje alterar alguma configuração do programa, basta selecionar a engrenagem situada mais à direita na barra de *menus*. É neste local, por exemplo, que se seleciona a exportação da matriz de resíduos do modelo e a opção de realizar o processamento dos regimes espaciais separadamente ou em conjunto.

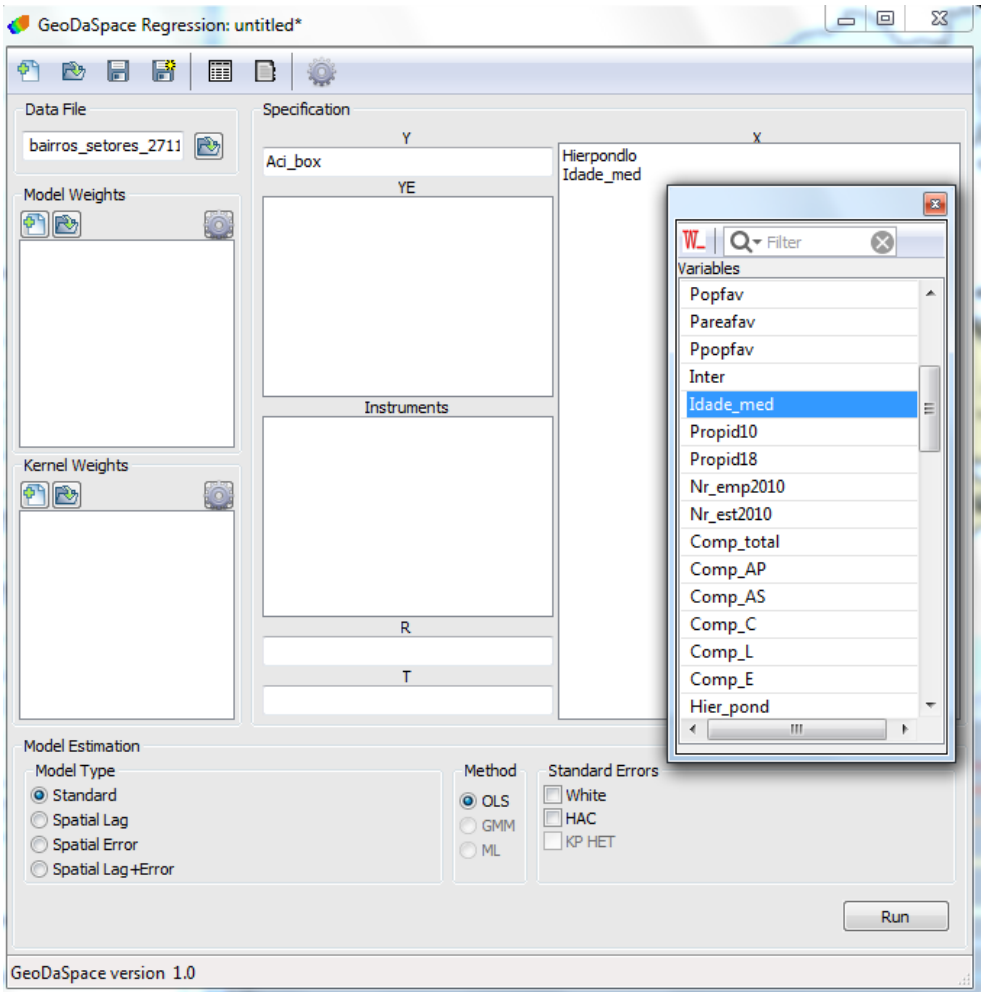


Figura 15 Tela do programa GeoDa Space

Além desses valores, outros diagnósticos podem ser utilizados para auxiliar na análise dos resultados da modelagem. A Figura 16 apresenta o sumário obtido no programa GeoDa Space. Pode-se observar em primeiro lugar o diagnóstico das variáveis explicativas. Costuma-se adotar o valor de 30 como sendo o limite máximo para o valor do número de condição de colinearidade. No caso deste programa, este também apresenta um teste de normalidade dos resíduos do modelo, o qual deve apresentar um resultado não significativo para que tenha uma distribuição próxima da normal.

Em seguida, deve-se verificar a existência ou não de dependência espacial, observando-se os valores dos índices de Moran e os multiplicadores de Lagrange. Para que estes valores sejam calculados é necessário antes especificar a matriz de proximidade.

```
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :bairros_setores_070115_residuos2.dbf
Weights matrix      :File: viz4_060115.gwt
Dependent Variable  :     Aci_box                Number of Observations:         119
Mean dependent var  :      2.0905                Number of Variables   :           3
S.D. dependent var  :      0.2819                Degrees of Freedom    :         116
R-squared           :      0.6864
Adjusted R-squared  :      0.6810
Sum squared residual:       2.942                F-statistic           :    126.9280
Sigma-square        :       0.025                Prob(F-statistic)     :   6.204e-30
S.E. of regression  :       0.157                Log likelihood        :      51.302
Sigma-square ML     :       0.025                Akaike info criterion :     -96.604
S.E of regression ML:      0.1572                Schwarz criterion     :     -88.266

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       0.2152265       0.1387574       1.5510995       0.1236019
          Hierpondlo       0.4057181       0.0411385       9.8622578       0.0000000
           Idade_med       0.0340672       0.0040117       8.4920062       0.0000000
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER           23.595

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2           1.504          0.4715

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                2           2.299          0.3169
Koenker-Bassett test              2           2.682          0.2616

SPECIFICATION ROBUST TEST
TEST                             DF        VALUE           PROB
White                             5           3.102          0.6842

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Moran's I (error)             0.1701         3.144          0.0017
Lagrange Multiplier (lag)          1         8.406          0.0037
Robust LM (lag)                    1         2.116          0.1458
Lagrange Multiplier (error)        1         7.717          0.0055
Robust LM (error)                  1         1.427          0.2323
Lagrange Multiplier (SARMA)        2         9.833          0.0073
```

Figura 16 Exemplo de sumário do programa GeoDa Space

Conforme mencionado anteriormente, o índice de Moran é um teste difuso de dependência espacial, podendo ser aplicado em conjunto com os testes mais focados, como é o caso dos multiplicadores de Lagrange. No entanto, as observações dos multiplicadores de Lagrange devem ser vistos como sendo apenas um indicador dos modelos que devem apresentar os melhores resultados. Primeiramente, deve-se observar os valores da significância dos multiplicadores de Lagrange. Para aqueles que apresentarem significância abaixo de 0,05 (para o caso do nível de confiança de 95%) deve-se verificar também a significância do teste robusto do modelo equivalente. Caso a significância esteja também abaixo de 0,05, este modelo poderá ser um bom modelo espacial.

Conforme mencionado anteriormente, a dependência espacial deve ser solucionada antes de se tratar da heterogeneidade espacial. Desse modo, os valores dos diagnósticos

apresentados antes da eliminação ou diminuição da dependência espacial (caso esteja presente) devem ser analisados em definitivo somente após contemplar esse efeito. A análise da heterogeneidade espacial será melhor vista no item seguinte.

### 4.8 Análise da heterogeneidade espacial

No programa GeoDa Space, os diagnósticos de heterogeneidade espacial são feitos observando-se os resultados dos resultados dos testes de Breusch-Pagan, Koenker-Bassett e White, sendo este último considerado o mais robusto deles, devendo, portanto, ser aquele adotado como o preferencial na análise da heterocedasticidade. Nestes testes, assim como no de normalidade, deve-se observar a significância do mesmo. Caso seja significativo, o modelo possui variâncias diferentes ao longo da região de trabalho, o que demandará em modelagens diferentes ao longo da região de trabalho.

Conforme mencionado na revisão bibliográfica, a heterogeneidade nos coeficientes pode ser feita de forma discreta ou contínua. Este trabalho empregará o modelo de regimes espaciais, o qual adota diferentes valores de coeficientes segundo um critério discreto. É comum na literatura obter os regimes espaciais obedecendo critérios tais como Norte e Sul, centro e periferia, ricos e pobres, etc., o que pode incorrer algumas vezes em uma certa dificuldade em se definir o limite destas sub-regiões. Como forma de evitar tal situação, empregou-se na divisão da região de estudo em regimes espaciais o método de regionalização REDCAP, o qual será melhor explicado no item sobre o MAUP. Este método necessita que se forneça uma variável, o número de divisões e o método de regionalização propriamente dito. A variável a ser empregada na regionalização será a variável dependente da modelagem. O número de divisões será função do número de polígonos que se esteja trabalhando. Não é recomendado trabalhar estatisticamente com regimes espaciais menores que 30 polígonos, o que acaba limitando o número de divisões da região de estudo. Por exemplo, caso se tenha 100 polígonos, seria recomendado que fosse feito, no máximo, uma divisão em três regimes. O programa mostra ainda a opção de se selecionar o mínimo de polígonos que se deseja ter em cada cluster. Caso se verifique que o programa esteja gerando clusters com menos de 30 polígonos, pode-se configurar o programa para um mínimo de 30. Dentre os métodos de regionalização, serão testados os métodos SLK, ALK e CLK, sendo o primeiro na opção Primeira Ordem (*First-Order*) e Todas as Ordens (*Full-Order*) e os demais somente na opção Todas as Ordens. A Figura 17 apresenta a tela do programa REDCAP. Na parte superior esquerda

são selecionados os dados sobre os quais são realizados o método de regionalização. Na parte central esquerda seleciona-se o método de regionalização, o número mínimo e máximo de polígonos que devem constar no cluster de saída (opcionalmente), os comandos para processar o modelo e salvar o resultado do modelo em formato CSV e *Shapefile*. Na parte inferior esquerda constam os resultados do processamento e na parte direita a parte gráfica com o mapa e as barras horizontais que funcionam como uma legenda que associa as cores do mapa com os valores numéricos dos dados de entrada.

Após cada seleção do número de divisões e o método utilizado, exporta-se o mesmo no formato CSV. Pode-se importar todos os arquivos CSV gerados no GeoDa Space em uma planilha eletrônica e exportar para o aplicativo de SIG para que sejam devidamente associado aos polígonos do nível geográfico utilizado na pesquisa.

No programa GeoDa Space, importa-se o dado geográfico com as informações geradas no REDCAP e processam-se os modelos considerado os regimes espaciais. Este programa permite processar os regimes espaciais separadamente ou em conjunto. É importante processar separadamente para se ter uma percepção do comportamento do modelo nas regiões separadamente. No entanto, o processamento em conjunto costuma apresentar melhores resultados que separadamente, embora este não possa ser descartado. O modelo a ser selecionado para esse teste é o melhor modelo selecionado na etapa de modelagem.

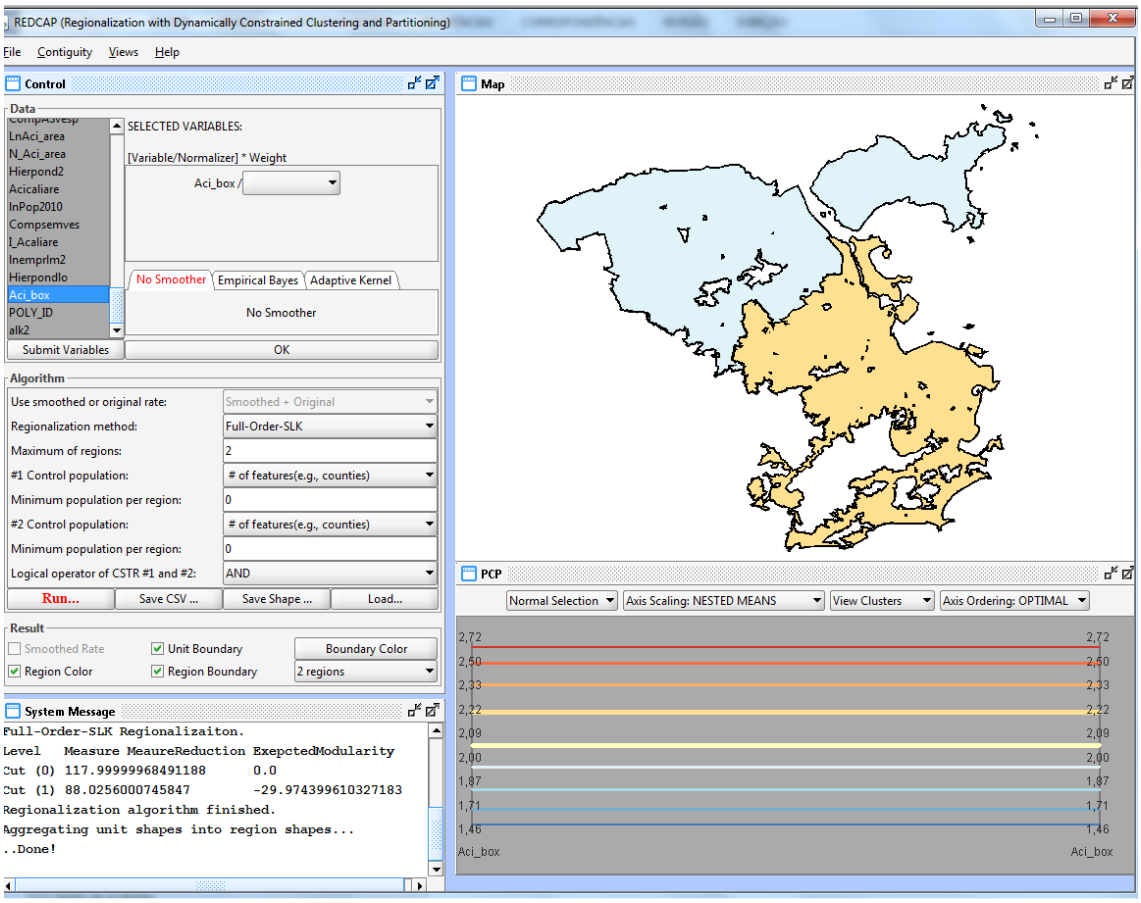


Figura 17 Tela do programa REDCAP

Após a seleção do melhor número de regimes espaciais e de método de regionalização, pode-se ainda fazer um outro teste que é o de processar o modelo levando em consideração a divisão em regimes espaciais, só que agora considerando o campo do regime espacial como mais uma variável explicativa. Quando se aplicam os modelos com regimes espaciais, é importante verificar o teste de Chow. Neste teste, deve-se observar a significância das variáveis. Aquelas cuja significância esteja abaixo do desejado (por exemplo, 5%) apresentam coeficientes diferentes para cada regime espacial. Falando em outros termos, um nível de significância acima de 5% mostra que os coeficientes são iguais em todos os regimes.

## 4.9 MAUP – Geração de dados

O problema do MAUP é uma das especificidades dos dados geográficos, o qual representa uma instabilidade nos resultados obtidos nas análises estatísticas dos dados agregados em área quando se muda o seu nível de agregação. Como forma de tentar lidar com esse problema e reduzir a falácia ecológica deverão ser buscadas regiões mais

homogêneas a partir da verificação do comportamento dos coeficientes das variáveis explicativas à medida que se modifica o nível de agregação.

Para tal, é necessário percorrer três etapas: definir qual o modelo a ser empregado na determinação dos coeficientes, o número de agregações que serão testadas e o critério a ser empregado na geração das diferentes áreas de agregação. O modelo a ser empregado deverá ser o melhor modelo selecionado nas etapas anteriores. O número de áreas de agregação deve variar entre um mínimo de 30 clusters, para que seja melhor tratado estatisticamente, e o valor máximo que coincidiria com o modelo gerado no item de modelagem. Neste intervalo, pode-se escolher o número de clusters que se julgue mais conveniente, não sendo recomendado um valor muito pequeno de intervalo entre os diferentes níveis de agregação, para que se possa melhor detectar as variações ocorridas nos modelos. Quanto ao critério para gerar as diferentes áreas de agregação, nesta metodologia será empregado o método de regionalização REDCAP, utilizado por XU *et al.* (2014) na comparação entre modelos de acidentes de trânsito.

O método de regionalização consiste em dividir um conjunto de objetos em um dado número de regiões contiguas, a partir da otimização de uma função objetivo, comumente uma medida de homogeneidade de uma dada grandeza presente nas regiões (GUO, 2008). A medida de homogeneidade no REDCAP é a soma total do quadrado dos desvios (GUO e WANG, 2011) representada pela Eq. 33.

$$SSD = \sum_{r=1}^{k} \sum_{i=1}^{n_r} \sum_{j=1}^{d} \left(x_{ij} - \bar{x}_j\right)^2 \qquad \text{Eq. 33}$$

Onde *k* é o número de regiões, $n_r$ é o número de objetos na região *r*, *d* é o número de variáveis consideradas, $x_{ij}$ é o valor do atributo *j* da região *i* e $\bar{x}_j$ é a média dos valores do atributo *j* para todos os objetos da região *r*. No caso desta metodologia, o valor de *d* será 1 tendo em vista que as regionalizações serão obtidas a partir da variável resposta. Dessa forma, serão obtidos um conjunto de k polígonos, onde se busca minimizar o valor do somatório do quadrado dos desvios.

Dentre os métodos de regionalização existentes, o REDCAP diferencia-se por ser um método que utiliza explicitamente a contiguidade das regiões. Por ser hierárquico, ao contrário do método de particionamento como o *k-means*, busca otimizar não a função objetivo global, mas faz o processo passo-a-passo encontrando em cada passo a melhor aglutinação de aglomerados vizinhos, o que não produz necessariamente a melhor função objetivo em níveis mais globais. Mais detalhes sobre os diferentes métodos de

regionalização podem ser vistos em GUO (2008) e GUO e WANG (2011).

O programa REDCAP é constituído por duas fases: a de agregação dos dados considerados contíguos de forma a produzir uma rede contínua espacialmente e a partição da rede de modo a se otimizar a função objetivo.

A primeira etapa leva em consideração critérios diferentes para a medição da distância entre os aglomerados, ou seja, a dissimilaridade entre os atributos da variável. São elas a ligação simples (S*ingle LinKage* - SLK), a qual considera a distância entre os aglomerados como sendo a menor distância entre os polígonos contidos em cada aglomerado, ligação média (*Average LinKage* – ALK), que faz a média entre os valores da distância entre cada um dos polígonos de um aglomerado com o dos demais polígonos do outro aglomerado e a ligação completa (*Complete LinKage* – CLK), que considera a distância entre os aglomerados como sendo a maior distância entre os polígonos contidos em cada aglomerado. Cada uma delas pode considerar somente a ligação entre os vizinhos como sendo de primeira ordem (*first-order*) ou de todas as ordens (*full-order*), totalizando um somatório de seis métodos.

Observando-se a Figura 18, os tons de cinza representam intensidades da variável adotada no estudo e as linhas representam as diferentes possibilidades de ligações entre os polígonos. Considerando-se dois aglomerados C1 (composto por A, B, C, D e E) e C2 (composto por F, G e H), as linhas cheias representam as ligações entre cada um dos polígonos e o aglomerado vizinho que compartilham lados (critério de primeira ordem) e as linhas tracejadas são as ligações que não compartilham lados. Neste caso o critério com todas as ordens seria o somatório das linhas cheias e tracejadas. Desse modo, os vizinhos de C1 pelo critério de primeira ordem seriam F e H e de C2 seriam B e E. No caso dos vizinhos de todas as ordens, os vizinhos de C1 seriam F, G e H e de C2 seriam A, B, C, D e E. O critério de medição de distância, por sua vez, obedecendo esta restrição de vizinhança, seria determinado obedecendo os valores de maior proximidade, maior distância ou pela média dos valores das distâncias entre os polígonos dos clusters C1 e C2 (Figura 18). Por exemplo, para o SLK de primeira ordem o valor de maior similaridade seria obtido entre os polígonos E e F e pelo SLK de todas as ordens seria o B e G por possuírem valores (cores) mais próximos.

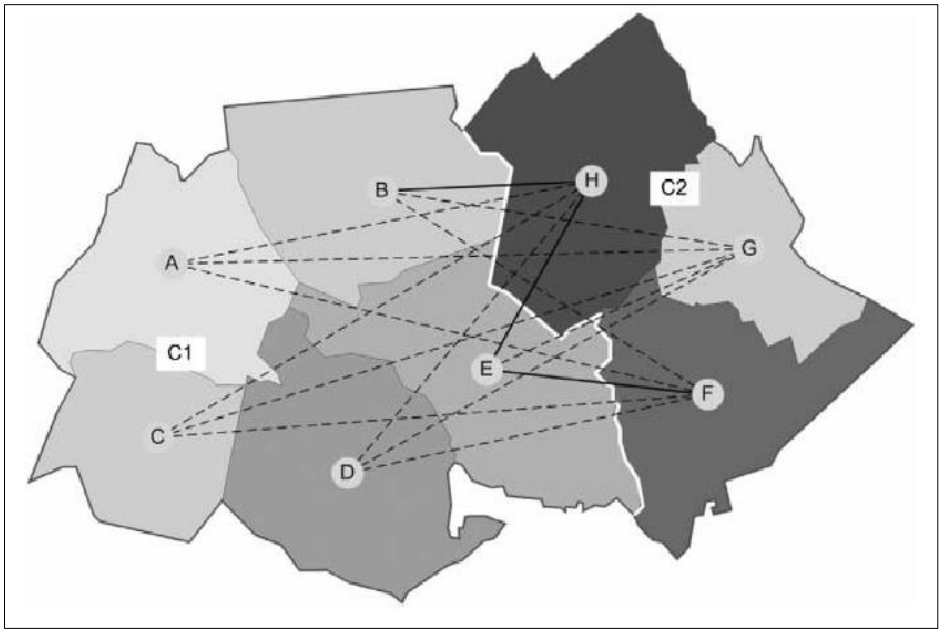


Figura 18 Medição de distâncias na etapa de criação de aglomerados
Fonte: GUO (2008)

Ao final da primeira etapa cada um dos polígonos estaria ligado a um outro polígono contíguo. Para o caso se querer determinar dois clusters, é preciso escolher qual a ligação deverá ser cortada para que se tenha os dois conjuntos de polígonos. Esta seria a segunda fase e utilizaria o critério da otimização da função objetivo, visto anteriormente, para determinar quais regiões seriam mais homogêneas entre si.

A saída do programa REDCAP é um arquivo texto com dois campos: um com o código do polígono e outro com um valor inteiro que o associa a um dado cluster. Por exemplo, para o caso de uma divisão em dois clusters, os polígonos poderão ter os valores 0 ou 1 de acordo com o fato de pertencerem a um ou outro cluster. No aplicativo de SIG pode-se facilmente associar este arquivo aos dados geográficos do nível de agregação dos polígonos. Empregando uma ferramenta que dissolva os polígonos baseando-se no valor dos clusters obtém-se os dados geográficos dos clusters. Neste momento, selecionam-se as variáveis mais explicativas, cujos valores serão somados quando da geração dos clusters. O cuidado que se precisa ter nesse momento é com os valores ponderados. Por exemplo, no caso da densidade de uma grandeza selecionam-se a área e a grandeza no momento de geração dos clusters. Após a dissolução recalculam-se a densidade a partir dos valores da área e da grandeza já agregados.

Por fim, aplicam-se os modelos selecionados na calibração sobre as variáveis calculadas após a agregação nos clusters e registram-se os valores dos coeficientes e do

seu erro padrão de cada um destes modelos. A análise de sensibilidade destes coeficientes será visto no item seguinte.

### 4.10 MAUP –Análise de sensibilidade

A análise de sensibilidade dos coeficientes dos modelos gerados no item anterior ocorrerá a partir da aplicação de um teste estatístico com a finalidade de verificar se os coeficientes dos modelos são considerados estatisticamente diferentes ou não.

A comparação ocorrerá entre os coeficientes da mesma variável explicativa dois a dois, entre aqueles situados em áreas de agregação adjacentes. Por exemplo, no caso de se ter os níveis de agregação 50, 60, 70, 80, 90 e 100, o coeficiente de uma dada variável no nível de agregação de 50 será comparado com 60, 60 será comparado com 70 e assim por diante.

O teste a ser empregado será o t-teste de Student para comparar a média de dois conjuntos de dados diferentes (agregações diferentes) e com variância da população desconhecida. As variâncias amostrais, por sua vez, podem ser iguais ou diferentes. Para esclarecer esta questão, deve-se aplicar um teste para variância como, por exemplo, o de Fisher-Snedecor.

Para o caso das variâncias serem iguais, deve-se aplicar o teste *t*, conforme a Eq. 34.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_P} \qquad \text{Eq. 34}$$

Sendo que o $S_p$ é dado pela Eq. 35.

$$S_p = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \qquad \text{Eq. 35}$$

Onde $s^2$ é um estimado imparcial da variância das duas amostras e $n_1$ e $n_2$ são as quantidades de participantes de cada conjunto de dados. Neste caso, a distribuição *t* possui $n_1 + n_2 - 2$ graus de liberdade.

Para o caso de variâncias serem diferentes, deve-se aplicar o teste *t* conforme a Eq.36, também conhecido como t-teste de Welch.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S} \qquad \text{Eq. 36}$$

Onde $S$ é dado pela Eq. 37.

$$S = \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}} \qquad \text{Eq. 37}$$

Para o caso do teste de significância, a distribuição do teste estatístico se aproxima de uma distribuição de *t* cujo graus de liberdade (Gl) são calculadas utilizando a Eq. 38.

$$Gl = \frac{\left({s_1}^2/n_1 + {s_2}^2/n_2\right)^2}{({s_1}^2/n_1)^2/(n_1 - 1) + \ ({s_2}^2/n_2)^2/(n_2 - 1)} \qquad \text{Eq. 38}$$

Além da verificar se os coeficientes são iguais dentro de um determinado nível de significância, pode-se também comparar visualmente os coeficientes por meio de um gráfico de dispersão conforme pode ser visto na Figura 19. Neste gráfico os pontos são valores do valor médio da variável e a linha vertical representa o erro padrão da mesma.

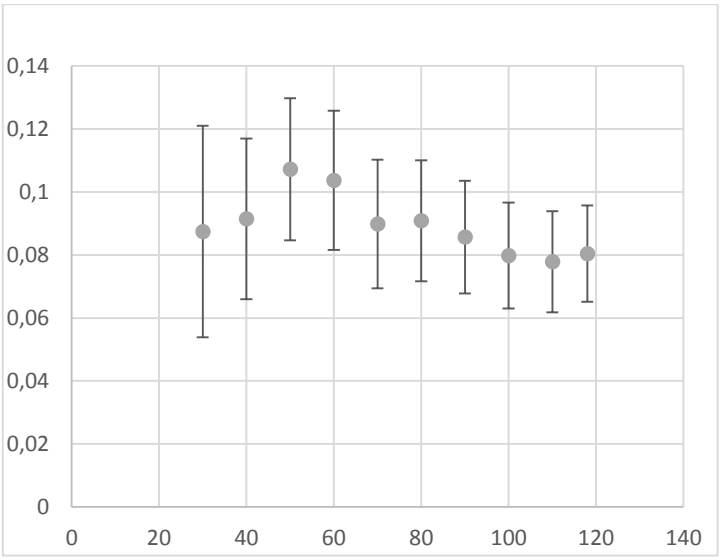


Figura 19 Gráfico com os valores dos coeficientes com barra vertical de erro

Na calibração, o tipo de gráfico apresentado na Figura 19 será utilizado não somente para observar os coeficientes das variáveis explicativas nos diferentes níveis de agregação, mas também o ajuste dos modelos e os valores da dependência e heterogeneidade espaciais dos resíduos, de modo a verificar se a mudança de agregação causa significativa mudança no comportamento das variáveis e, consequentemente, dos resultados dos modelos.

### 4.11 Validação dos modelos

Dentre as formas de validação possíveis, esta pesquisa construirá o intervalo de previsão para a variável dependente no modelo de validação e verificará se o valor

observado encontra-se dentro do intervalo de previsão do modelo. Serão também calculadas outras grandezas como o coeficiente de Pearson e o NRMS, conforme visto da revisão bibliográfica, de forma a comparar os validar os parâmetros determinados na calibração.

Tendo em vista que a metodologia emprega as variáveis como sendo dados geográficos, é importante que na etapa de validação sejam feitas as análises visual e explicativa das variáveis envolvidas na validação. Assim como na calibração, é importante verificar se a questão do MAUP se aplica aos dados da validação, assim como os efeitos da dependência e heterogeneidade espaciais.

Desse modo, a validação não deve se resumir somente ao cálculo de estatísticas, mas passa também pela melhor compreensão dos dados trabalhados nesta etapa.

## 5. ANÁLISE DE RESULTADOS

### 5.1 Caracterização da área de estudo

O Rio de Janeiro é uma cidade com cerca de 6,4 milhões de habitantes, segundo o Censo Demográfico de 2010 produzido pelo IBGE, estando dividida em três grandes regiões: zona Norte, zona Sul e zona Oeste. Tem como característica peculiar a de estar às margens do oceano Atlântico e da baía da Guanabara e de possuir grandes extensões de área verde composta por remanescentes florestais de mata atlântica e com relevo montanhoso. Na Figura 20, pode-se observar em destaque três grandes áreas verdes: o maciço da Tijuca mais à direita, o maciço da Pedra Branca ao centro e o maciço do Mendanha mais ao norte. Quanto às grandes massas d´água, à leste tem-se a baía da Guanabara, onde está situada a ilha do Governador, na parte inferior o oceano Atlântico e à oeste a baía de Sepetiba.

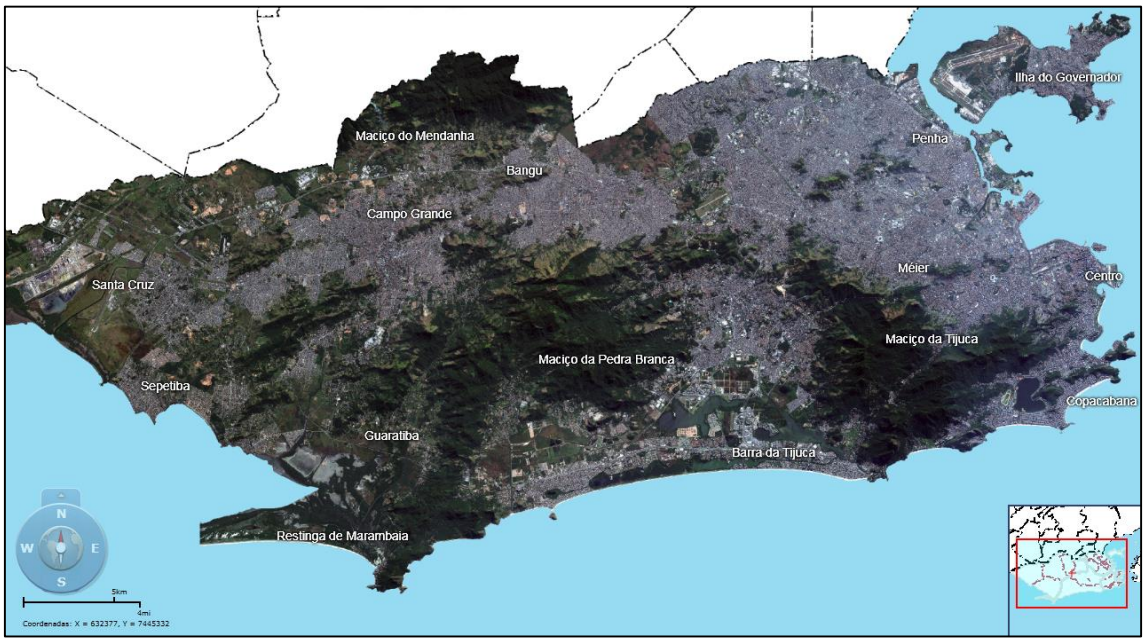


Figura 20 Imagem de satélite da cidade do Rio de Janeiro destacando áreas verdes e massas d´água
Fonte: SIG Floresta (2012).

Adotou-se como área de trabalho as zonas Norte e Sul da cidade, incluindo as ilhas do Fundão e do Governador. A ilha de Paquetá foi eliminada da pesquisa por ter pequeno fluxo de veículos, não sendo registrados acidentes de trânsito no período estudado. Na parte mais a leste do município, entre as zonas Norte e Sul, está situado o Centro (*Central Businness District* - CBD) do município.

Embora apresentem condições socioeconômicas diferentes, as zonas Norte e Sul apresentam como característica comum entre si e com o centro da cidade a de possuírem centros de alcance metropolitano para aonde confluem grande quantidade de pessoas da cidade do Rio de Janeiro e dos municípios vizinhos, seja para trabalhar, seja em busca de lazer, comércio e serviços. Tal fato tem um forte impacto nos acidentes de trânsito, como será detalhado mais à frente.

A área adotada como a menor unidade geográfica na pesquisa foi a de bairros, podendo ser agregados em regiões maiores para se obter regiões mais homogêneas no que diz respeito aos acidentes de trânsito. Não se adotou a unidade geográfica dos setores censitários pelo fato dos mesmos, na região em estudo, serem muito pequenos (média de 0,44 $km^2$) para que sejam obtidas as variáveis explicativas empregadas no estudo. A região apresenta 119 bairros (desconsiderando a ilha de Paquetá). O limite dos bairros foi obtido a partir da união de 6351 setores censitários do censo de 2010 do IBGE. Destes setores, 122 apresentaram valores zerados e foram, portanto, eliminados. Dos 122 setores eliminados, 60 estavam contidos em regiões com cota acima de 100 metros, regiões estas comumente cobertas por área verde. A Figura 21 apresenta os bairros empregados na pesquisa, identificando-se aqueles pertencentes às nonas Norte, Sul e Central. Os espaços em branco no meio dos mesmos representam os setores censitários eliminados conforme mencionado anteriormente.

Comparando-se os mapas contidos nas Figuras 20 e 21, é possível perceber que a zona Sul está delimitada pelo oceano Atlântico ao sul, pelo Parque Nacional da Tijuca ao norte e oeste e pela baía de Guanabara a leste. A zona Norte está limitada ao sul e oeste pelo Parque Nacional da Tijuca, pela baía da Guanabara a leste e outros municípios ao norte. Nesta região também estão contidas as principais ilhas habitáveis do município. São elas: a ilha do Fundão, a ilha do Governador e a ilha de Paquetá. A zona Oeste está separada geograficamente das demais zonas pelo Parque Nacional da Tijuca e Parque Estadual da Pedra Branca.

O Rio de Janeiro, por sua vez, é um município com grande extensão de área verde, como pode ser visto na Figura 20, sendo que grande parte destas áreas verdes estão situadas em áreas montanhosas. Para se ter uma ideia da extensão das áreas verdes na região de trabalho, dos 6351 setores censitários, 297 apresentam centroide dentro de uma região cuja cota está acima de 100 metros. Mesmo representando menos de 5% do número total de setores censitários, estes respondem por mais de 20% em área de trabalho.

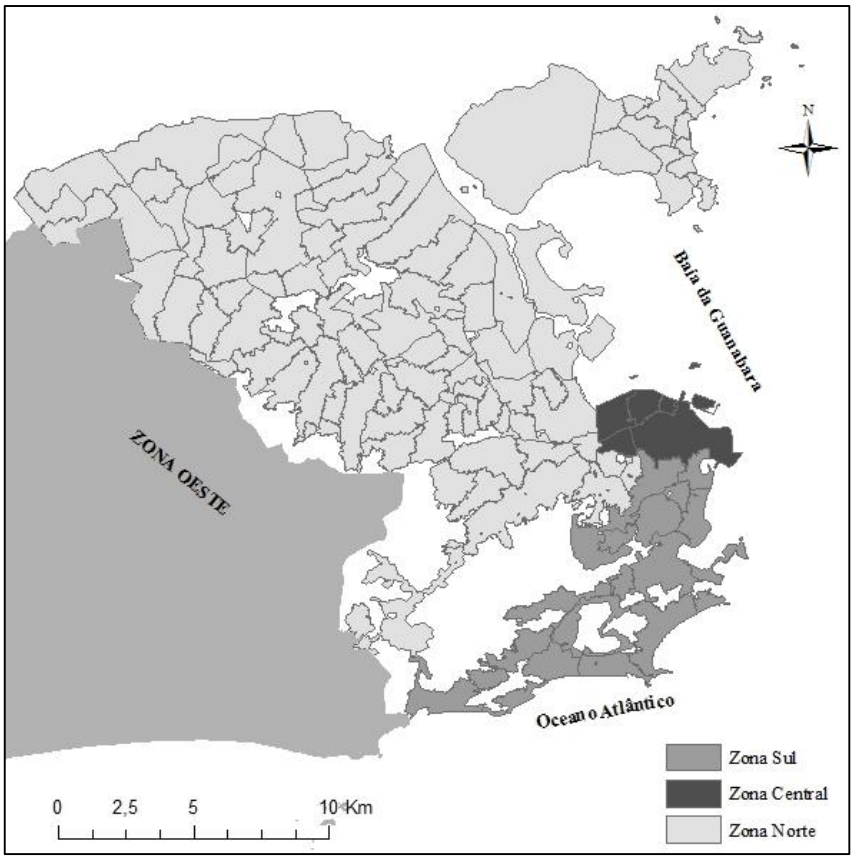


Figura 21 Bairros do Rio de Janeiro empregados na pesquisa

Outra característica da cidade do Rio de Janeiro é a presença de grande quantidade de aglomerados subnormais, também chamados popularmente de favelas. Dos 6351 setores censitários, 1366 são classificados pelo IBGE como sendo aglomerados subnormais. Mesmo representando cerca de 22% dos setores censitários, ocupam menos de 10% da área, por serem regiões densamente povoadas (próximo de 50000 hab/km$^2$ para uma densidade média da área de trabalho de cerca de 35000 hab/km$^2$). Na Figura 22 é possível observar a grande quantidade de aglomerados subnormais da região de estudo, principalmente na região da zona Norte.

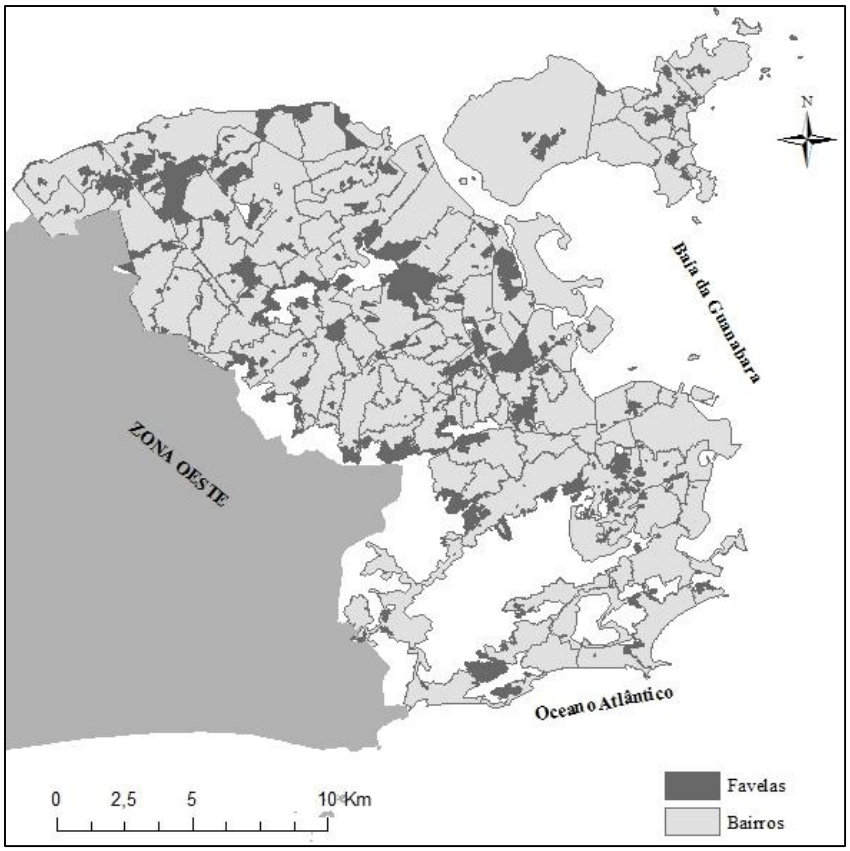

Figura 22 Aglomerados subnormais (favelas) da região de estudo

## 5.2 Preparação dos dados

Após a etapa de seleção das variáveis a partir da caracterização da área de estudo e da verificação da bibliografia sobre o assunto, inicia-se a preparação dos dados para as etapas de compreensão da distribuição espacial e de modelagem dos dados. Os programas computacionais a serem empregados na pesquisa a partir da preparação dos dados são os seguintes:

- aplicativo de Sistema de Informações Geográficas ArcGIS 10;
- programas de análise exploratória de dados geográficos GeoDa e de Econometria Espacial GeoDa Space;
- programas estatísticos R e SPSS 15.

### 5.2.1 Dados de acidentes

Os dados de acidentes em grande parte das vias foram obtidos pela Polícia Militar do Rio de Janeiro no local do acidente com o emprego de rastreadores de satélites *Global Positioning System* (GPS) e fornecidos na forma de coordenadas X,Y em uma planilha

eletrônica. A Tabela 6 apresenta o número de acidentes nos anos de 2008 a 2011, coletados pela Polícia Militar do Rio Janeiro, sendo que os dados de 2008 a 2010 serão empregados para construir os modelos e os de 2011 para a validação dos mesmos. Nesta tabela também constam a frota de veículos no município do Rio de Janeiro considerando todos os tipos para ilustrar a relação entre o aumento dos acidentes e da frota. Estes dados não foram empregados no estudo por estarem disponibilizados somente no nível do município e não de bairros.

Tabela 6 Frota de veículos e número de acidentes no município do Rio de Janeiro nos anos de 2008 a 2011

| Ano | Frota | Variação (%) | Nr de acidentes | Variação (%) |
|---|---|---|---|---|
| **2008** | 1841274 | - | 45763 | - |
| **2009** | 1947622 | 5,78 | 44611 | - 2,51 |
| **2010** | 2063521 | 5,95 | 48013 | 7,63 |
| **2011** | 2190395 | 6,14 | 53263 | 10,93 |

É importante observar que houve um aumento nos anos de 2010 e 2011 dos acidentes acima do aumento da frota de veículos de todos os tipos. No entanto, foi abaixo do aumento do número de motocicletas, cujo aumento foi em torno de 11% para todos os anos. A diminuição do número de acidentes coincide com o início da chamada Operação Lei Seca no Estado do Rio de Janeiro, que se iniciou em 19 de março de 2009, o qual contribuiu para a redução no número de acidentes no Rio de Janeiro naquele ano.

A Polícia Militar não forneceu os dados de algumas vias administradas pela CET-Rio denominadas de vias especiais, vias estas de grande importância para o trânsito da cidade e que servem como que de espinha dorsal da rede viária do município. A prefeitura da cidade, por meio da CET-Rio, monitora com mais detalhamento estas vias e atende os veículos em pane e os acidentes ocorridos na mesma. Além das vias especiais, outras vias como a Linha Amarela e a Via Dutra não tiveram seus dados fornecidos, possivelmente pelo fato de serem administrados por concessionárias, onde os atendimentos aos usuários são feitos por elas mesmas. A descrição das vias especiais consideradas na pesquisa e da Linha Amarela constam da Tabela 7. Aqui não está mencionada a Via Dutra por constar somente em um pequeno trecho da área de estudo.

Tabela 7 Descrição das vias não utilizadas na pesquisa

| Vias | Descrição das vias |
|---|---|
| Linha Vermelha | Campo de São Cristóvão (via elevada) até a Via Dutra. |
| Avenida Brasil | Rodoviária Novo Rio até a estrada do Camboatá (altura de Deodoro da Av. Brasil). |
| Praça XV | Praia de Botafogo até o terminal rodoviário Novo Rio. |
| Rebouças | Viaduto Saint Hilaire até o campo de São Cristóvão, passando pela Av. Eng. Freyssinet e Eng. Rufino de Almeida Pizarro (vias elevadas). |
| Santa Bárbara | Largo de Santo Cristo até a praia de Botafogo, passando pelo viaduto Santiago Dantas em um sentido e pelas ruas Muniz Barreto e Voluntários da Pátria no outro. |
| Autoestrada Lagoa-Barra | Av. Min. Ivan Lins até a avenida Borges de Medeiros na altura do clube do Flamengo. |
| Linha Amarela | Ilha do Fundão até praça do pedágio em Água Santa. |

Verificou-se que todas as vias especiais, com exceção de parte da via especial Praça XV e a via especial Santa Bárbara, são classificadas pela prefeitura do Rio de Janeiro como sendo vias estruturais. Estas vias especiais classificadas como estruturais coincidem em grande parte com as rodovias federais e estaduais que cortam o município. Arbitrou-se que tais vias seriam retiradas da modelagem, tendo em vista o fato dos dados de acidentes estarem disponíveis somente em formato de planilha eletrônica e agregados por ano, o que levaria a uma grande imprecisão da estimativa dos acidentes por trecho de via. Somado a isso, grande parte destas vias são segregadas e predominantemente utilizadas como vias de passagem. As Figuras 23 e 24 mostram exemplos de vias excluídas da pesquisa. A Figura 23 mostra as duas pistas da Linha Vermelha em níveis diferentes sobre a rua lateral ao Campo de São Cristóvão, no bairro de São Cristóvão. Como se pode observar, a realidade do bairro pouco tem a ver com o fluxo de carros que cruzam o bairro nesta via expressa. A Figura 24 mostra as pistas da avenida Infante Dom Henrique, situada no Aterro do Flamengo, quase que totalmente segregadas do bairro do Flamengo, que se encontra ao fundo da imagem. Apesar de serem utilizadas por moradores dos bairros em que cruzam, o seu uso é bem superior como via de passagem.


Figura 23 Via lateral do Campo de São Cristóvão
Fonte: *Google Maps*


Figura 24 Avenida Infante Dom Henrique
Fonte: *Google Maps*

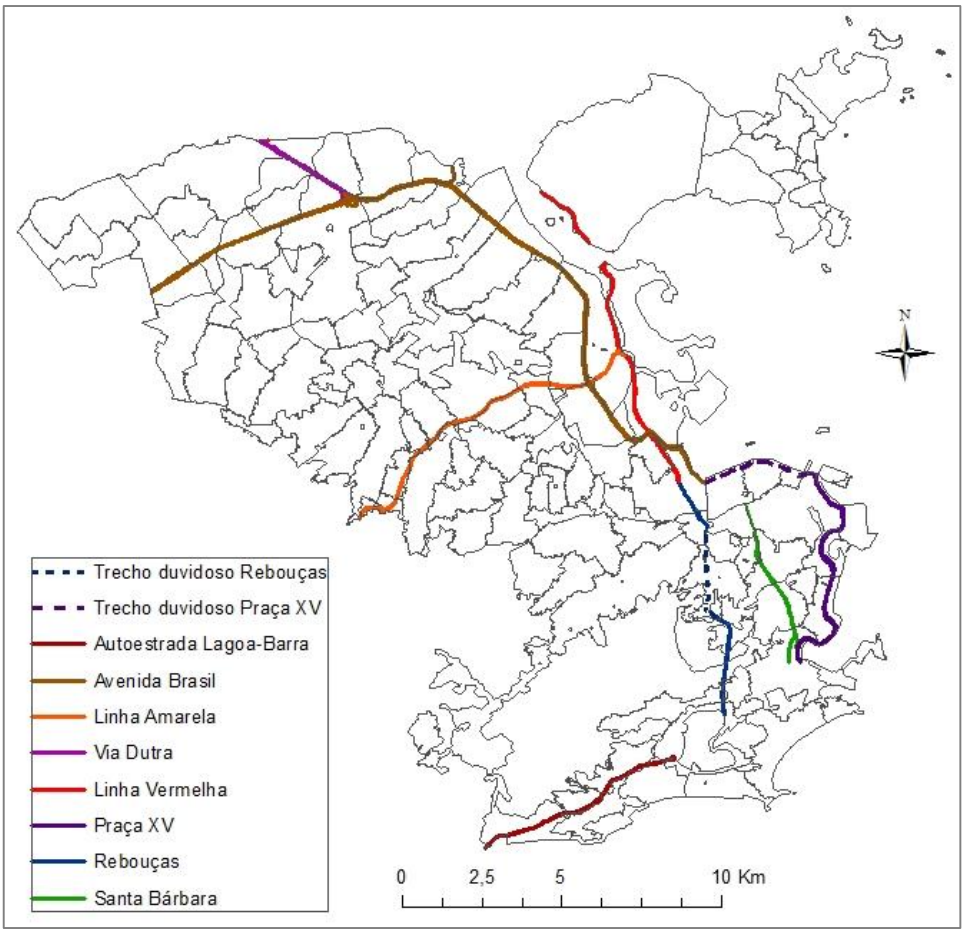


Figura 25 Vias não utilizadas na pesquisa

Mantiveram-se somente os trechos de vias estruturais da Av. Borges de Medeiros e Av. Francisco Bicalho por serem vias estruturais não segregadas e terem seus dados de acidentes da PM reportados, ao contrário dos demais trechos de vias estruturais que não tiveram acidentes reportados. A Figura 25 apresenta um mapa contendo as vias excluídas da pesquisa.

Dentre as vias especiais que são elevadas, duas delas geraram dúvida se os acidentes fornecidos pela Polícia Militar pertenciam à via especial ou à avenida situada abaixo desta, mesmo sendo descritas com os nomes das vias situadas abaixo (Figura 25). Tal dúvida surgiu devido ao fato das avenidas situadas abaixo possuírem traçado paralelo à via especial e também grande fluxo de veículos. São elas parte do elevado Engenheiro Freyssinet que se encontra sobre a avenida Paulo de Frontin (via especial Túnel Rebouças) e parte da avenida Presidente Juscelino Kubitschek situada sobre a avenida Rodrigues Alves (via especial Praça XV). Para retirar tais dúvidas, analisou-se o número de acidentes por km das vias interligadas com as mesmas de forma a verificar se a quantidade de acidentes era coerente com o da via em análise. Analisou-se também o

padrão dos locais onde estavam concentrados os acidentes, a partir da verificação dos locais com maior concentração de interseções. Por outro lado, observaram-se também os trechos das vias próximos onde havia concentração de acidentes para verificar a existência de algum padrão que justificasse tal disposição. Verificou-se que os acidentes estavam situados próximos das interseções nas vias situadas abaixo das vias especiais e que os trechos nestas vias tinham o traçado muitas vezes reto o que não justificaria tal concentração de acidentes. Por isso e pela coerência do número de acidentes com o das vias do entorno, considerou-se os acidentes integralmente como pertencendo às vias situadas abaixo das vias especiais. Verificou-se também que os trechos das vias especiais antes e depois destes trechos elevados não apresentavam registros pela Polícia, o que reforça essa ideia.

Considerou-se na pesquisa os acidentes nos anos de 2008, 2009 e 2010, de modo a atenuar as flutuações dos valores de acidentes de um ano para outro. Além disso, a quantidade de acidentes é dividida pelo valor da área do bairro para eliminar o viés de se ter valores maiores de acidentes em uma dada área de agregação simplesmente pelo fato desta apresentar o maior valor de área. Dessa forma, a variável explicativa a ser utilizada é a média da densidade dos acidentes nos anos de 2008, 2009 e 2010. Por questão de simplificação, será referida somente como densidade de acidentes.

Outra variável possível de ser empregada é a relação entre a quantidade de acidentes e a somatório da extensão de todas as vias na área de agregação, obtendo um valor de acidentes por km de via. Ao se verificar a correlação entre a área dos bairros e a extensão das vias, obteve-se o valor de 0,76, indicando uma forte correlação, provavelmente devido ao fato do limite dos bairros ser obtido a partir da agregação dos setores censitários cujo valor é diferente de zero, ou seja, não estarem consideradas regiões do bairro sem população, tais como, áreas verdes e massas d'água. No entanto, tendo em vista a tese trabalhar com variáveis agregadas em áreas, preferiu-se empregar a variável densidade de acidentes (acidentes / $km^2$).

É importante, quando se preparam as variáveis a serem empregadas na modelagem, tentar reduzir a variância das mesmas, de modo a tornar a ordem de grandeza das mesmas mais próximas das demais variáveis. A aplicação do logaritmo (caso seja uma relação adequada entre a variável explicativa e resposta) ou simplesmente a mudança da unidade da variável podem ajudar nesse sentido.

Para se ter uma ideia, a ordem de grandeza da hierarquia encontra-se entre 1 e 5, a idade média entre 28 e 45 anos, enquanto que a população possui média em torno de

40000 e os empregos por $km^2$ uma média de 5500. Caso fossem deixados dessa forma, os coeficientes calculados para as variáveis com maiores valores apresentariam valores muito pequenos. Dentre as variáveis utilizadas na pesquisa, aplicou-se o logaritmo sobre as variáveis população e número de empregos e adotou-se a unidade de $ha^2$ no lugar de $km^2$ para a densidade da população mais emprego.

### 5.2.2 Variáveis associadas à geometria das vias

As variáveis associadas à geometria das vias públicas foram obtidas a partir da base cartográfica da cidade do Rio de Janeiro, na escala de 1:10000, produzida pelo Instituto Pereira Passos, pertencente à prefeitura da cidade do Rio de Janeiro. São elas:

- extensão das vias dentro da área de agregação;
- hierarquia das vias ponderada pela extensão das mesmas em cada área de agregação.

Cabe aqui ressaltar que as hierarquias das vias, por serem variáveis qualitativas, não poderão ser empregadas diretamente nos modelos estatísticos, sendo convertidas em variáveis quantitativas a partir da atribuição de pesos às mesmas. Esta abordagem que utiliza diretamente a hierarquia das vias traduzida em pesos não foi encontrada na revisão bibliográfica. Quando muito, encontrou-se a extensão ou percentual de vias de uma região dentro de cada hierarquia. Por questões de simplificação, tal variável será referida no texto somente como hierarquia ponderada. Os pesos adotados encontram-se na Tabela 8.

Tabela 8 Pesos empregados na determinação da variável hierarquia ponderada

| **Hierarquias** | **Pesos** |
|---|---|
| Estruturais | 5 |
| Arteriais primárias | 4 |
| Arteriais secundárias | 3 |
| Coletoras | 2 |
| Locais | 1 |

Espera-se que a hierarquia ponderada funcione como uma alternativa ao fluxo de veículos comumente empregada na modelagem de acidentes, ou mesmo à largura das vias por ser estar relacionada ao fluxo de veículos.

### 5.2.3 Variáveis associadas à conectividade das vias

As informações de conectividade das ruas a serem consideradas são:

- número de interseções dentro da área de agregação;

- densidade de interseções nas áreas de agregação.

Cabe aqui ressaltar que a diferença entre o número de nós e o de interseções está no número de pontos que representam o final dos segmentos de ruas, este contabilizado somente no número de nós.

### 5.2.4 Variáveis demográficas

As variáveis demográficas foram obtidas do censo demográfico do IBGE de 2010 e são as seguintes:

- densidade populacional, obtida pela razão entre o somatório da população na área de agregação e o valor da área dessa região;

- percentual da população das favelas do bairro em relação à população do bairro;

- percentual da área de favelas do bairro em relação à área do bairro;

- idade média da população em cada área de agregação, calculada conforme a Eq. 37.

$$I_{med} = \frac{\sum_{0,5}^{99} n.Pop_n}{\sum_{0,5}^{99} Pop_n} \qquad \text{Eq. 37}$$

Onde *n* representa o valor da idade. Neste caso, considerou-se a idade de menos de um ano com peso 0,5. O numerador contém o somatório do produto entre o valor da idade e a população com essa idade e o denominador o somatório de todas as idades, ou seja, a população total de cada área de agregação.

As variáveis que fazem referência às favelas não foram encontradas nos estudos referidos na revisão bibliográfica, por estes ocorrerem comumente em países em desenvolvimento como o Brasil.

A idade também não costuma ser utilizada na forma de idade média e sim na forma de contagem ou percentual da população em determinadas faixas etárias. A idade média da população foi testada como forma de retratar uma realidade presente no Rio de Janeiro de que os bairros mais ricos possuem um maior percentual de idosos fruto de uma maior expectativa de vida aliado a um menor número de nascimentos, ao contrário do que ocorre nos bairros mais pobres, principalmente naqueles com maior quantidade de pessoas residentes em favelas. Essa variável tem a característica de ser constituída por pessoas de

todas as idades, refletindo no resultado a maior ou menor influência de cada uma das idades.

### 5.2.5 Variáveis socioeconômicas

As variáveis socioeconômicas empregadas na pesquisa foram as seguintes:

- renda per capita, calculada pela razão entre o somatório da renda de todos os moradores dos domicílios acima de 10 anos e o número de habitantes de cada região de agregação, obtida do Censo 2010;

- número absoluto e densidade de estabelecimentos, fornecido no Relatório Anual do Trabalho e do Emprego fornecido pelo Ministério do Trabalho e Emprego (RAIS/MTE) referente ao ano de 2012 e obtido no site Armazém Digital da Prefeitura do município do Rio de Janeiro;

- número absoluto e densidade de empregos com carteira assinada, fornecido pelo (RAIS/MTE), referente ao ano de 2012;

- a densidade do somatório da população e do emprego. Para fins de simplificação, será referida no texto como densidade da população mais emprego.

Esta última variável não foi encontrada de forma conjunta na referência bibliográfica. Foi inserida como forma de retratar, em uma mesma variável duas variáveis muito empregadas na modelagem, mas que ao mesmo tempo não sejam muito correlacionadas com as demais variáveis explicativas.

O nível de agregação da renda é o setor censitário, enquanto que o do número de estabelecimentos e a quantidade de emprego são os bairros.

### 5.2.6 Variáveis associadas à acessibilidade aos transportes públicos

As variáveis associadas à acessibilidade aos transportes públicos foram fornecidas pela Federação das Empresas de Transportes de Passageiros do Estado do Rio de Janeiro (FETRANSPOR). São elas:

- quantidade de pontos de ônibus; e

- quantidade de linhas de ônibus. Cada linha de ônibus pode ser considerada duas vezes, caso a linha passe na área de agregação no trajeto de ida e no trajeto de volta.

Esta última variável foi empregada como forma de contemplar uma característica da cidade do Rio de Janeiro que é a dos transportes públicos estarem concentrados no modo

rodoviário e haver grande quantidade de linhas de ônibus.

### 5.3 Verificação dos efeitos de bordas dos acidentes

Para a verificação dos efeitos de bordas, construiu-se *buffer zones* de largura 15 metros na borda de todos os polígonos de bairros, ou seja, 15 metros para ambos os lados da borda dos polígonos. Tal largura é um valor arbitrado, baseado na precisão da coordenada do sistema GPS na época em torno de 8 metros e a largura das faixas de rolamento em torno de 3 metros. Considerando-se que grande parte das vias possui entre duas e três faixas de rolamento o que somado com a incerteza de 8 metros do sistema GPS, fornece um valor próximo de 15 metros para cada um dos lados. Em alguns programas, como é o caso do ArcGIS, é necessário converter o polígono em linhas para que se obtenha, a partir destas linhas de contorno, as *buffer zones*. Obtiveram-se para todos os três anos (2008, 2009 e 2010) um total de cerca de 5% do total de acidentes no interior destas regiões, incluindo os acidentes ocorridos nos bairros vizinhos e que não constam na área de estudo. Tal valor julgou-se ser pequeno em relação ao total de acidentes, o que não justificaria uma abordagem que contemplasse separadamente os acidentes ocorridos no interior dos bairros e na região das bordas dos mesmos.

### 5.4 Análise visual

Em um primeiro momento, fez-se a análise visual dos acidentes a partir da construção de mapas coropléticos, a partir de três métodos: iguais valores, quebra natural e desvio padrão. Conforme mencionado, o critério de quebra natural coloca na mesma classe da legenda valores mais próximos entre si a partir da minimização das variâncias das classes, ou seja, no caso do mapa com quatro classes as legendas contemplam as quatro menores variâncias para este conjunto de dados. O dos iguais valores divide a legenda em intervalos de mesmo tamanho e o mapa de desvio padrão mostra as classes com valores múltiplos de desvio padrão.

O mapa de quebra natural da densidade dos acidentes nos anos de 2008 a 2010 pode ser visto na Figura 26. É possível verificar a existência de maior densidade de acidentes na região em torno do Centro e da zona Sul da cidade, diminuindo quando se vai na direção norte.

O mapa de iguais valores da mesma variável pode ser visto na Figura 27. Pode-se

observar que existem quatro bairros (4/119) que constam na classe com maiores valores de acidentes/área, 9/119 constam na segunda classe mais alta, 23/119 na terceira classe mais alta e 83/119 (cerca de 70%) dos bairros constam na classe de menores valores. Tal fato mostra a concentração de bairros com maior quantidade de acidentes em poucos bairros.

O mapa de desvio-padrão da média dos acidentes nos anos de 2008 a 2010, dividido pela área pode ser visto na Figura 28. Pode-se observar com desvio padrão superior a 2,5 desvios padrão os mesmos quatro bairros que constam na classe com maiores valores de densidade de acidentes no mapa de iguais valores. Um valor próximo do número de bairros dessa classe também consta entre 1,5 e 2,5 desvios padrão. Considerando que a média é de cerca de 187, 51/119 valores estão em torno da média (entre -0,5 e 0,5 desvios padrão) e 41/119 valores estão abaixo de 0,5 desvios padrão. Tal fato mostra que o valor da média é baixo em relação ao intervalo que se atinge os acidentes. Mais detalhes serão vistos no item sobre análise exploratória.

Tendo em vista que se tem dados pontuais de acidentes, gerou-se o mapa de densidade de Kernel para se tenha uma ideia da distribuição dos acidentes dentro dos bairros. Quando se superpõe os dados das vias sobre o mapa de densidade de Kernel, é possível verificar em torno de quais vias há concentração de acidentes. A Figura 29 apresenta um mapa de densidade de Kernel gerado a partir dos dados de acidentes do ano de 2009 superposto pelas vias estruturais e arteriais. É possível verificar que grande parte das regiões com maior densidade de acidentes são cortadas por tais vias.

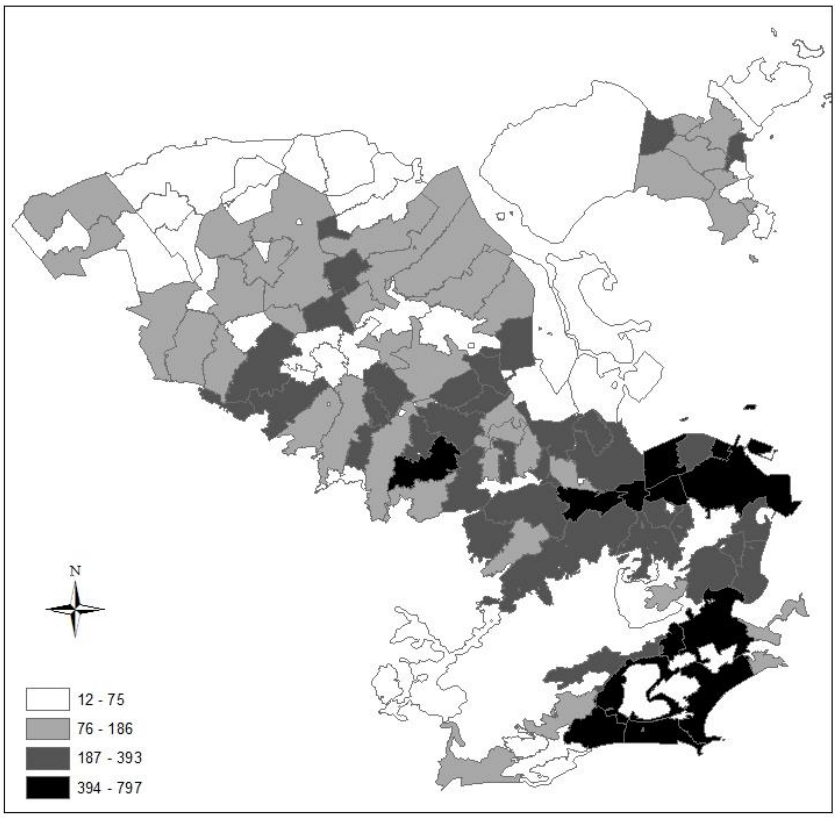


Figura 26 Densidade média de acidentes pelo critério da quebra natural

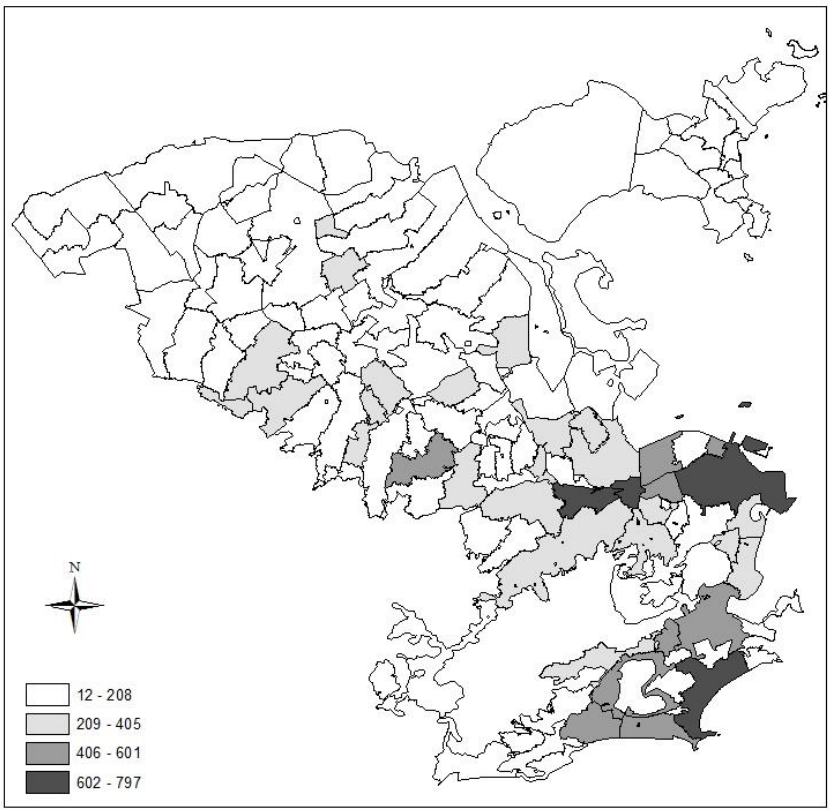


Figura 27 Densidade média de acidentes pelo critério dos iguais valores

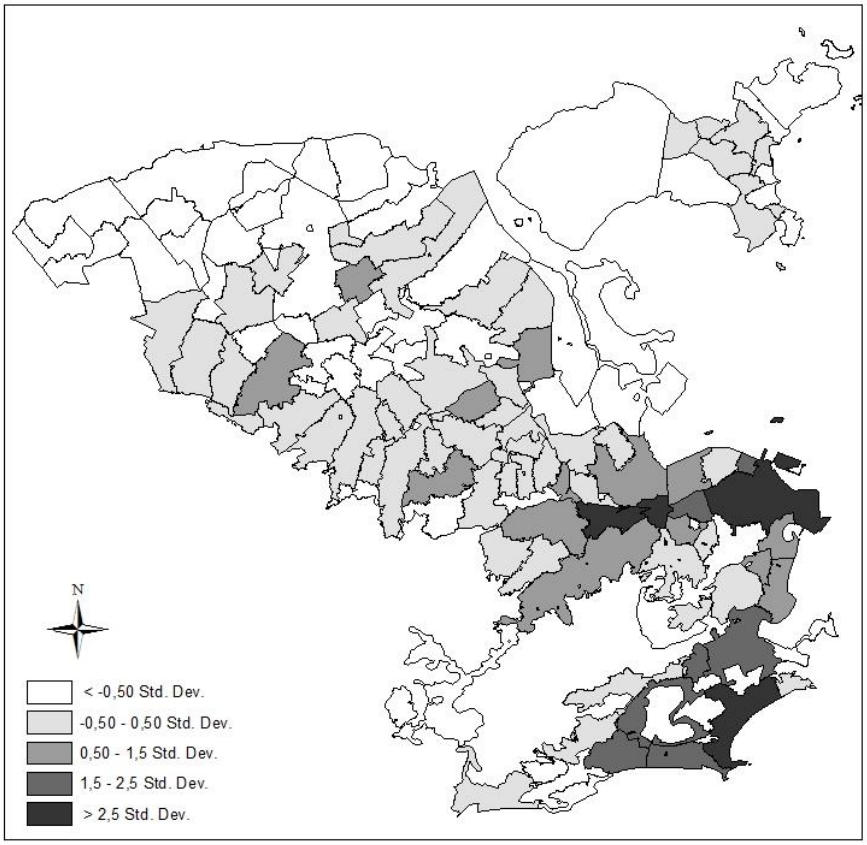


Figura 28 Densidade média de acidentes pelo critério do desvio padrão

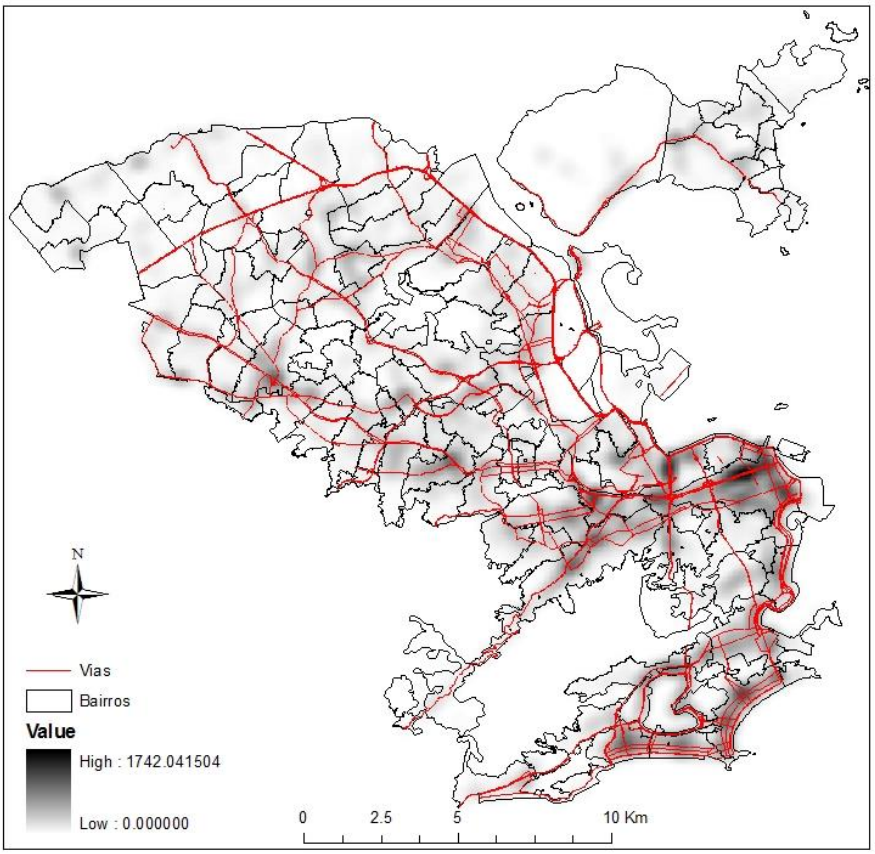


Figura 29 Mapa de densidade de Kernel com as vias estruturais e arteriais

## 5.5 Análise exploratória

A análise exploratória concentrou-se inicialmente sobre a variável resposta, cuja distribuição foi observada em um diagrama de caixa ou *boxplot*. Neste gráfico destacam-se cinco valores: o primeiro quartil (ou quartil inferior) situado na parte inferior da caixa; a mediana, representada pela linha horizontal no interior da caixa; o terceiro quartil (ou quartil superior) localizado na parte superior da caixa; *outlier* inferior representado pela linha horizontal abaixo da caixa e *outlier* superior representado pela linha horizontal acima da caixa. O valor do *outlier* inferior é obtido diminuindo do quartil inferior o intervalo interquartílico multiplicado por 1,5 ou 3. O valor do *outlier* superior é obtido somando ao quartil superior o intervalo interquartílico multiplicado por 1,5 ou 3. O intervalo interquartílico, por sua vez, é a diferença entre os quartis superior e inferior do diagrama de caixas. A visualização deste diagrama pode ser feita espacialmente por meio do *box map*. As Figuras 30 e 31 apresentam o *boxplot* e seu respectivo *box map* para os valores de 1,5 e 3,0 vezes o intervalo interquartílico, respectivamente. Observando-se o *boxplot* com valores de 1,5 vezes o intervalo interquartílico, é possível verificar a existência de 11 *outliers* superiores. No respectivo *box map* é possível verificar que os bairros considerados *outliers* superiores são os bairros da zona Sul, do Centro e próximos ao Centro. Quando se muda para 3,0 vezes o intervalo interquartílico, verifica-se que somente dois bairros continuam, sendo um deles o Centro da cidade.

Passando para a determinação da matriz de vizinhança, é importante ter em mente que a definição desta matriz influenciará sobremaneira os valores da autocorrelação espacial, bem como o cálculo da dependência espacial e consequentemente os modelos de regressão espacial.

Observando os mapas dos bairros, é possível perceber que os bairros apresentam tamanhos e número de vizinhos muito diferentes. Na região norte tem-se bairros pequenos com maior número de vizinhos e na região mais ao sul estão aqueles com maior tamanho e com poucos vizinhos. Além disso, tem-se uma parte continental e outra de ilhas, o que tende a diminuir a quantidade de vizinhos.

Dentre os critérios de vizinhança mais empregados, aqueles que adotam uma vizinhança física, como é o caso dos critérios da torre e da rainha, tendem a dificultar o caso das ilhas e dos bairros com poucos vizinhos que façam fronteira física. Adotar um critério de raio fixo pode trazer dificuldades na definição deste raio, tendo em vista que

para a região com bairros menores há uma tendência de se adotar um número de vizinhos muito maior que nos bairros com maior raio. Por outro lado, é desejável que se tenha um valor mínimo de vizinhos que possa fazer com que não se fique excessivamente dependente de poucos vizinhos. Por isso, selecionou-se o critério de um valor fixo de vizinhos mais próximos. De forma a facilitar a definição do número de vizinhos, construiu-se a matriz de vizinhança pelo critério da rainha, de modo a verificar o número de vizinhos físicos de cada um dos bairros. Obteve-se o valor médio de 4,49 vizinhos, dos quais cerca de 53% dos bairros apresentavam um número de vizinhos menor ou igual a 4 vizinhos e 73% apresentavam um número de vizinhos menor ou igual a 5 vizinhos. De forma a se obter um valor de número de vizinhos mais próximo da média do número de vizinhos adotou-se o valor dos 4 vizinhos mais próximos.

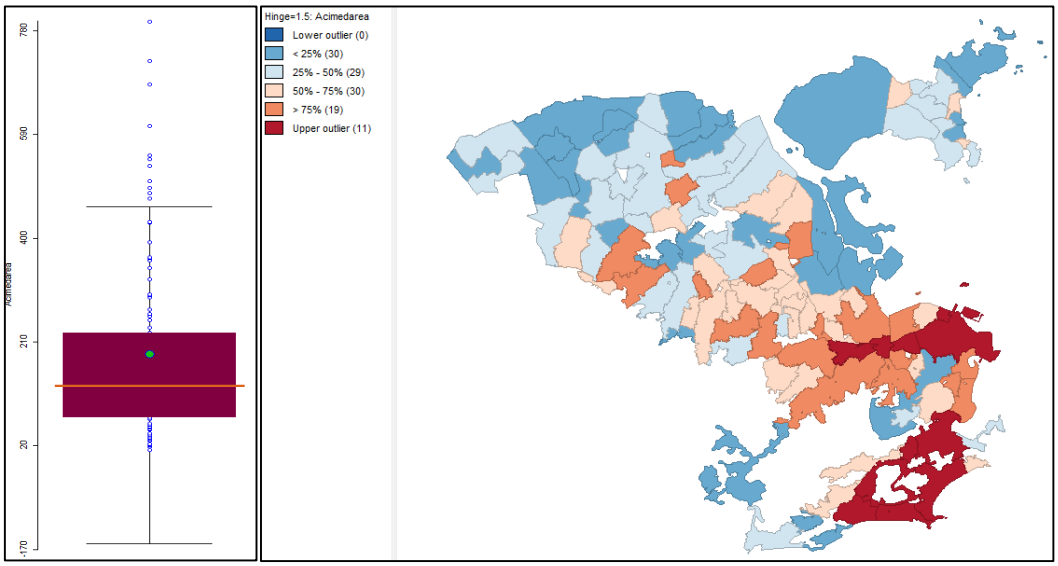


Figura 30 *Box plot* e respectivo *box map* para um valor de 1,5 vezes o intervalo interquartílico

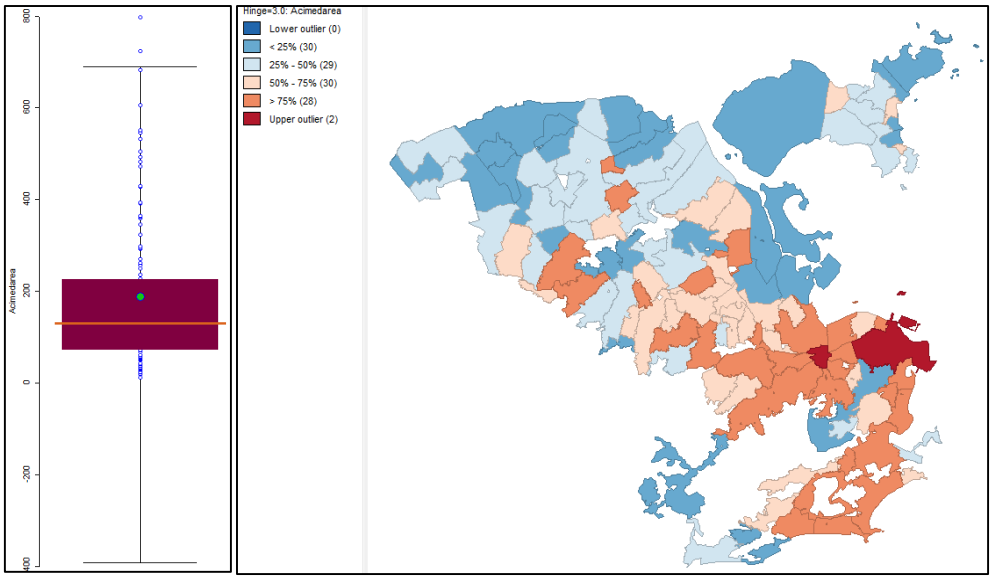


Figura 31 *Box plot* e respectivo *box map* para um valor de 3 vezes o intervalo interquartílico

Como forma de detectar a heterogeneidade espacial, construiu-se o diagrama de dispersão de Moran com o mapa de desvio padrão para verificar a existência de regimes espaciais com critério de vizinhança pelos 4 vizinhos mais próximos (Figura 32). Selecionando-se os bairros do quadrante Alto-Alto, observa-se que os bairros destacados são preponderantemente os das regiões mais ao centro e sul da região de trabalho. Os bairros da região mais ao norte seriam selecionados caso fossem selecionados os bairros do quadrante Baixo-Baixo, ou seja, aqueles bairros valores abaixo da média dos acidentes, cuja média móvel também se encontra abaixo dos vizinhos.

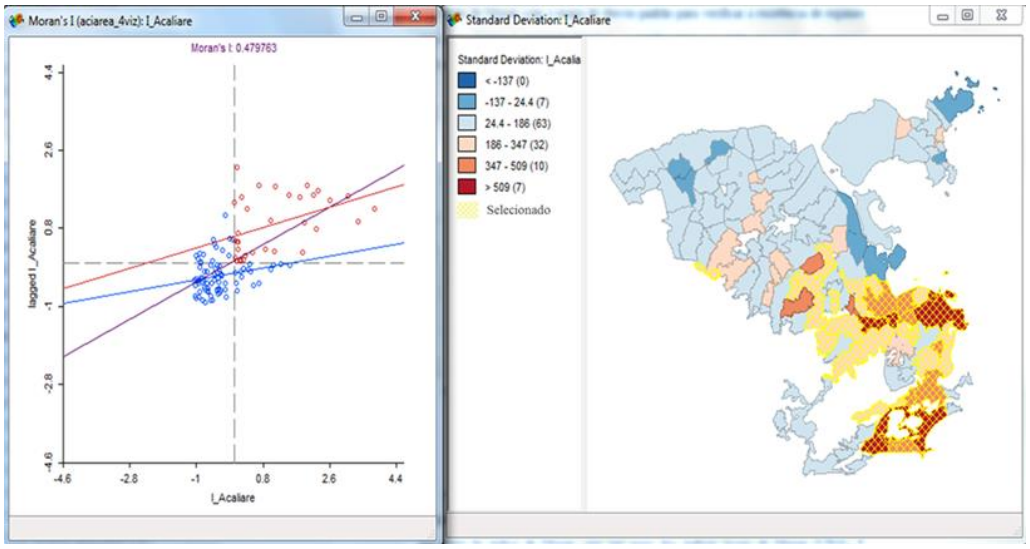


Figura 32 Diagrama de dispersão e mapa de desvio padrão para os 4 vizinhos mais próximos

Como forma de explorar os dados na busca de melhor identificar os limites dos possíveis regimes espaciais empregou-se um raio que pudesse incluir um maior número de vizinhos. A Figura 33 mostra o mapa LISA para um raio de 10 km.

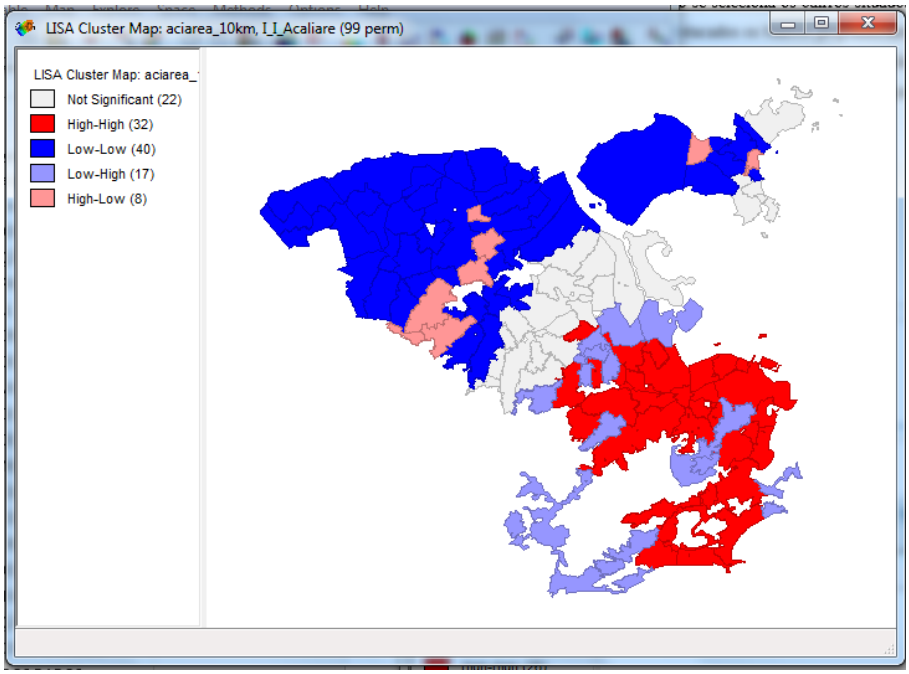


Figura 33 Diagrama de dispersão e mapa de desvio padrão para raio de 10 km

A partir da observação da Figura 33 é possível verificar a existência de duas regiões bem distintas que comportam valores de acidentes mais homogêneas entre si e que poderão vir a originar dois regimes espaciais.

Outra forma de observar a possibilidade de existência de regimes espaciais seria por meio da ANOVA espacial. Ao se utilizar como variável *dummy* o campo referente a dois regimes obtido do programa REDCAP, é possível verificar o quão a variável *dummy* é significativa e que somente ela contribui para um valor de $R^2$ajustado em torno de 0,37 (Figura 34).

```
Data set            :bairros_setores_271114_sem_centro_residuos.dbf
Weights matrix      :File: 4viz_space.gwt
Dependent Variable  :     Aci_box                Number of Observations:         118
Mean dependent var  :      2.0855                Number of Variables   :           2
S.D. dependent var  :      0.2777                Degrees of Freedom    :         116
R-squared           :      0.3762
Adjusted R-squared  :      0.3708
Sum squared residual:       5.628                F-statistic           :     69.9596
Sigma-square        :       0.049                Prob(F-statistic)     :    1.54e-13
S.E. of regression  :       0.220                Log likelihood        :      12.094
Sigma-square ML     :       0.048                Akaike info criterion :     -20.188
S.E of regression ML:      0.2184                Schwarz criterion     :     -14.647

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       2.2639591       0.0294354      76.9128959       0.0000000
          red2alkful      -0.3396552       0.0406083      -8.3641853       0.0000000
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER            2.504

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2           1.110          0.5742

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                1           0.064          0.8008
Koenker-Bassett test              1           0.073          0.7877

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Lagrange Multiplier (lag)         1          10.624          0.0011
Robust LM (lag)                   1           1.722          0.1894
Lagrange Multiplier (error)       1          16.887          0.0000
Robust LM (error)                 1           7.985          0.0047
Lagrange Multiplier (SARMA)       2          18.609          0.0001
```

Figura 34 Resultado da ANOVA espacial

Quando se aplica a superfície de tendência, é possível verificar que ambas as variáveis contribuem para um $R^2$ ajustado em torno de 0,26 (Figura 35). Ambas as variáveis também apresentam grande significância e possuem sinais dos coeficientes diferentes, sendo que o eixo X apresenta o sinal positivo mostrando que os acidentes crescem na direção Oeste para Leste. O eixo Y, por sua vez, apresenta o sinal negativo indicando um aumento Norte para o Sul. É importante lembrar que as coordenadas estão no sistema de coordenadas UTM, cujas coordenadas X aumentam sempre para o leste, independente do hemisfério. No caso das coordenadas Y, no hemisfério sul as coordenadas diminuem seus valores na direção sul.

```
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :bairros_setores_271114_sem_centro_residuos.dbf
Dependent Variable  :     Aci_box                Number of Observations:         118
Mean dependent var  :      2.0855                Number of Variables   :           3
S.D. dependent var  :      0.2777                Degrees of Freedom    :         115
R-squared           :      0.2688
Adjusted R-squared  :      0.2561
Sum squared residual:       6.597                F-statistic           :     21.1393
Sigma-square        :       0.057                Prob(F-statistic)     :   1.519e-08
S.E. of regression  :       0.240                Log likelihood        :       2.722
Sigma-square ML     :       0.056                Akaike info criterion :       0.557
S.E of regression ML:      0.2365                Schwarz criterion     :       8.869

White Standard Errors
------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT     155.0205759      37.4660316       4.1376300       0.0000671
              X_cent       0.0000077       0.0000034       2.2514450       0.0262569
              Y_cent      -0.0000212       0.0000049      -4.3349587       0.0000314
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER         3425.858

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2           5.349          0.0689

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                2           6.696          0.0351
Koenker-Bassett test              2           6.550          0.0378
```

Figura 35 Resultado da aplicação da superfície de tendência

## 5.6 Seleção das variáveis explicativas

A seleção das variáveis explicativas foi feita aplicando-se o coeficiente de correlação de Pearson (Tabela 9). Procurou-se analisar primeiramente a correlação das variáveis com a variável dependente e, num segundo momento, a correlação entre as variáveis explicativas, retirando-se aquelas que tenham correlação acima de 0,4 com as variáveis explicativas mais correlacionadas.

As variáveis explicativas que apresentaram maior correlação com a variável dependente foram: a hierarquia ponderada, a idade média e a densidade de estabelecimentos. No entanto, ao se verificar que a densidade de estabelecimentos possui uma correlação acima de 0,4 com ambas as variáveis e menor com a variável resposta que as outras duas variáveis supracitadas, não foi possível mantê-la no modelo.

Tabela 9 Coeficiente de Pearson entre as variáveis empregadas na pesquisa

| | Densidade de acidentes | Hierarquia ponderada | Comp vias | Nr interseções | Densidade interseções | Idade média | Renda média | Densidade demográfica | Log População | % área favelas | % pop favelas | Log densi empregos | Densidade estab | Densidade pop mais empregos | Nr pontos ônibus | Nr linhas ônibus | Coord X | Coord Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Densidade de acidentes | 1,000 | ,704 | -,060 | -,059 | ,129 | ,656 | ,430 | ,117 | ,033 | -,398 | -,508 | ,579 | ,635 | ,482 | ,265 | ,411 | ,327 | -,491 |
| Hierarquia ponderada | ,704 | 1,000 | -,090 | -,144 | -,059 | ,351 | ,389 | ,016 | -,068 | -,193 | -,191 | ,402 | ,431 | ,347 | ,133 | ,409 | ,381 | -,535 |
| Comp vias | -,060 | -,090 | 1,000 | ,953 | ,232 | -,030 | -,077 | -,109 | ,626 | -,118 | -,091 | ,130 | ,164 | ,092 | ,762 | ,449 | -,273 | ,185 |
| Nr interseções | -,059 | -,144 | ,953 | 1,000 | ,372 | -,058 | -,127 | -,051 | ,611 | -,058 | -,088 | ,109 | ,163 | ,137 | ,714 | ,426 | -,330 | ,247 |
| Densidade interseções | ,129 | -,059 | ,232 | ,372 | 1,000 | ,054 | -,226 | ,056 | ,143 | -,099 | -,229 | ,078 | ,092 | ,109 | ,139 | ,166 | -,260 | ,327 |
| Idade média | ,656 | ,351 | -,030 | -,058 | ,054 | 1,000 | ,711 | ,041 | -,008 | -,616 | -,759 | ,485 | ,571 | ,234 | ,236 | ,145 | ,321 | -,389 |
| Renda média | ,430 | ,389 | -,077 | -,127 | -,226 | ,711 | 1,000 | ,043 | -,044 | -,347 | -,409 | ,284 | ,420 | ,109 | ,066 | -,006 | ,427 | -,591 |
| Densidade demográfica | ,117 | ,016 | -,109 | -,051 | ,056 | ,041 | ,043 | 1,000 | ,470 | ,508 | ,236 | -,258 | ,196 | ,595 | ,009 | -,100 | ,061 | -,291 |
| Log População | ,033 | -,068 | ,626 | ,611 | ,143 | -,008 | -,044 | ,470 | 1,000 | ,258 | ,153 | -,047 | ,195 | ,308 | ,589 | ,264 | -,282 | -,053 |
| % área favelas | -,398 | -,193 | -,118 | -,058 | -,099 | -,616 | -,347 | ,508 | ,258 | 1,000 | ,810 | -,521 | -,249 | ,175 | -,154 | -,120 | -,164 | -,026 |
| % pop favelas | -,508 | -,191 | -,091 | -,088 | -,229 | -,759 | -,409 | ,236 | ,153 | ,810 | 1,000 | -,524 | -,334 | -,023 | -,231 | -,123 | -,056 | ,083 |
| Log densidade empregos | ,579 | ,402 | ,130 | ,109 | ,078 | ,485 | ,284 | -,258 | -,047 | -,521 | -,524 | 1,000 | ,518 | ,224 | ,284 | ,397 | ,233 | -,220 |
| Densidade estab | ,635 | ,431 | ,164 | ,163 | ,092 | ,571 | ,420 | ,196 | ,195 | -,249 | -,334 | ,518 | 1,000 | ,751 | ,506 | ,548 | ,343 | -,393 |
| Densi pop + empregos | ,482 | ,347 | ,092 | ,137 | ,109 | ,234 | ,109 | ,595 | ,308 | ,175 | -,023 | ,224 | ,751 | 1,000 | ,413 | ,464 | ,268 | -,373 |
| Nr pontos ônibus | ,265 | ,133 | ,762 | ,714 | ,139 | ,236 | ,066 | ,009 | ,589 | -,154 | -,231 | ,284 | ,506 | ,413 | 1,000 | ,624 | -,136 | -,105 |
| Nr linhas ônibus | ,411 | ,409 | ,449 | ,426 | ,166 | ,145 | -,006 | -,100 | ,264 | -,120 | -,123 | ,397 | ,548 | ,464 | ,624 | 1,000 | ,073 | -,132 |
| Coord X | ,327 | ,381 | -,273 | -,330 | -,260 | ,321 | ,427 | ,061 | -,282 | -,164 | -,056 | ,233 | ,343 | ,268 | -,136 | ,073 | 1,000 | -,314 |
| Coord Y | -,491 | -,535 | ,185 | ,247 | ,327 | -,389 | -,591 | -,291 | -,053 | -,026 | ,083 | -,220 | -,393 | -,373 | -,105 | -,132 | -,314 | 1,000 |

Quando se observou que as demais variáveis ligadas à geometria e conectividade, verificou-se que a extensão das vias, o número de interseções e a densidade de interseções apresentaram baixa correlação com a variável resposta. A fato da variável hierarquia ponderada apresentar maior correlação em relação ao somatório da extensão das vias mostra uma maior influência das vias de maior hierarquia em relação às demais hierarquias de vias.

Quanto às variáveis demográficas, a idade média foi a que apresentou o melhor resultado, embora a renda média tenha apresentado também bons resultados, sendo eliminada por ter forte correlação com a idade média.

Em se tratando das variáveis socioeconômicas, a densidade de empregos, de estabelecimentos e do somatório da população com os empregos obtiveram todas uma boa correlação com os acidentes. No entanto, somente a última apresentou uma correlação com as variáveis hierarquia ponderada e idade média abaixo de 0,4 e por isso foi selecionada. Testou-se também se o logaritmo do somatório da população com os empregos. No entanto, embora se obtivesse bons resultados na modelagem, a mesma foi considerada colinear com as outras duas variáveis. Estas variáveis foram testadas não somente na forma de densidade mas também em valores absolutos, obtendo-se em todos os casos valores inferiores aos da densidade. Os percentuais da área e da população em favelas apresentaram boa correlação com a variável resposta. No entanto, foram muito correlacionadas com a idade média, sendo por isso eliminadas.

Os pontos de ônibus e linhas de ônibus, embora sejam correlacionados com a densidade de acidentes, foram retirados pois apresentavam forte correlação com a densidade de empregos e de estabelecimentos.

Em resumo, as variáveis selecionadas nesta etapa foram a hierarquia ponderada, a idade média e a densidade da população mais empregos, cujo sumário estatístico, juntamente com a da variável resposta encontram-se na Tabela 10. A variável densidade da população mais empregos empregará a unidade hectares ao quadrado para que ordem de grandeza desta variável fique mais próxima da grandeza das demais explicativas e, em consequência, obtenha-se coeficientes também da mesma ordem de grandeza. Os mapas da variável dependente e das variáveis selecionadas pelo critério do desvio padrão encontram-se na Figuras 36 a 39.

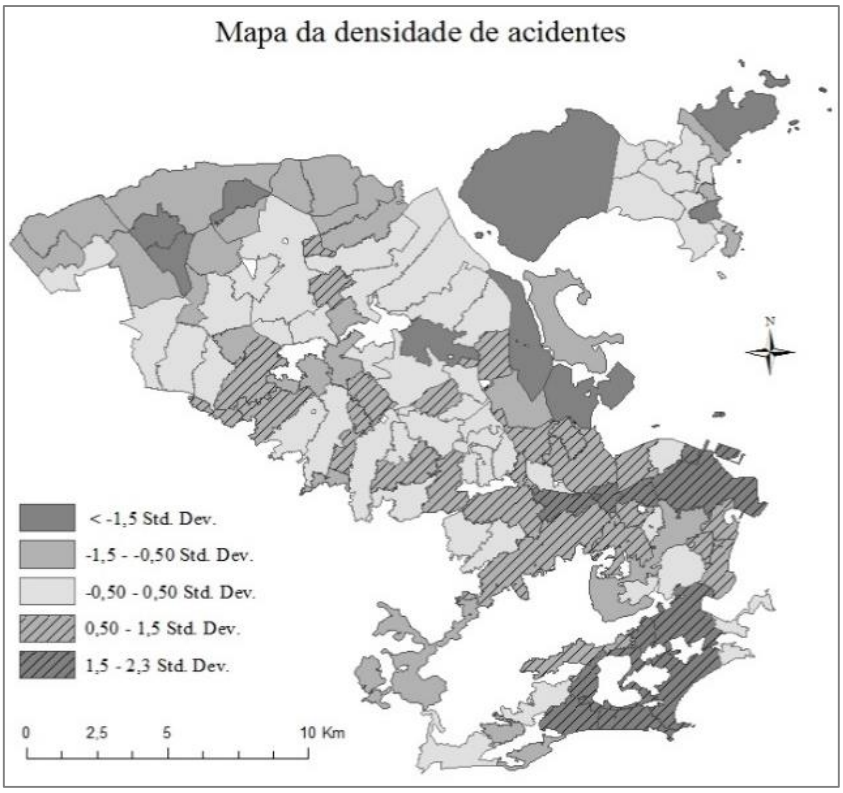


Figura 36 Mapas da variável densidade de acidentes

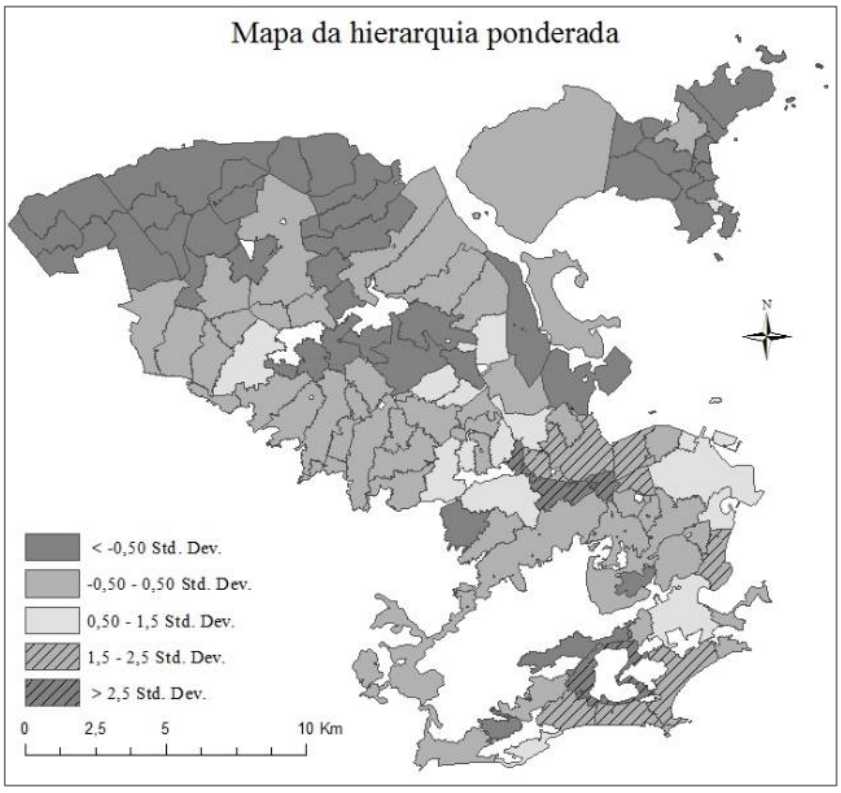


Figura 37 Mapa da variável hierarquia ponderada

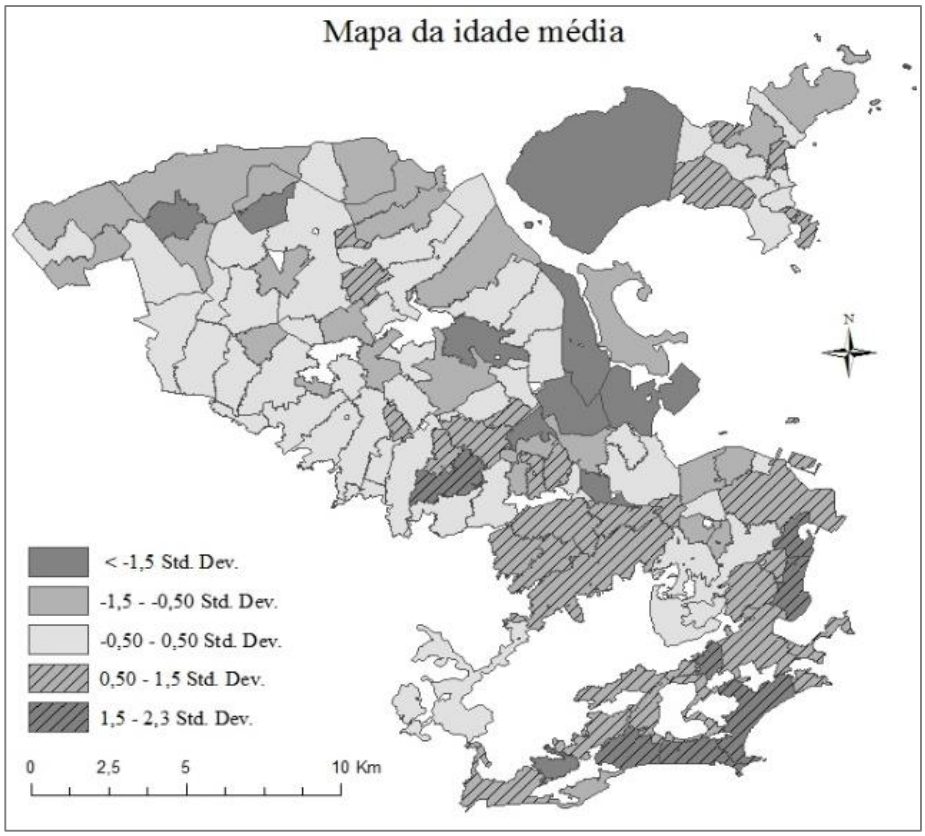


Figura 38 Mapa da variável idade média

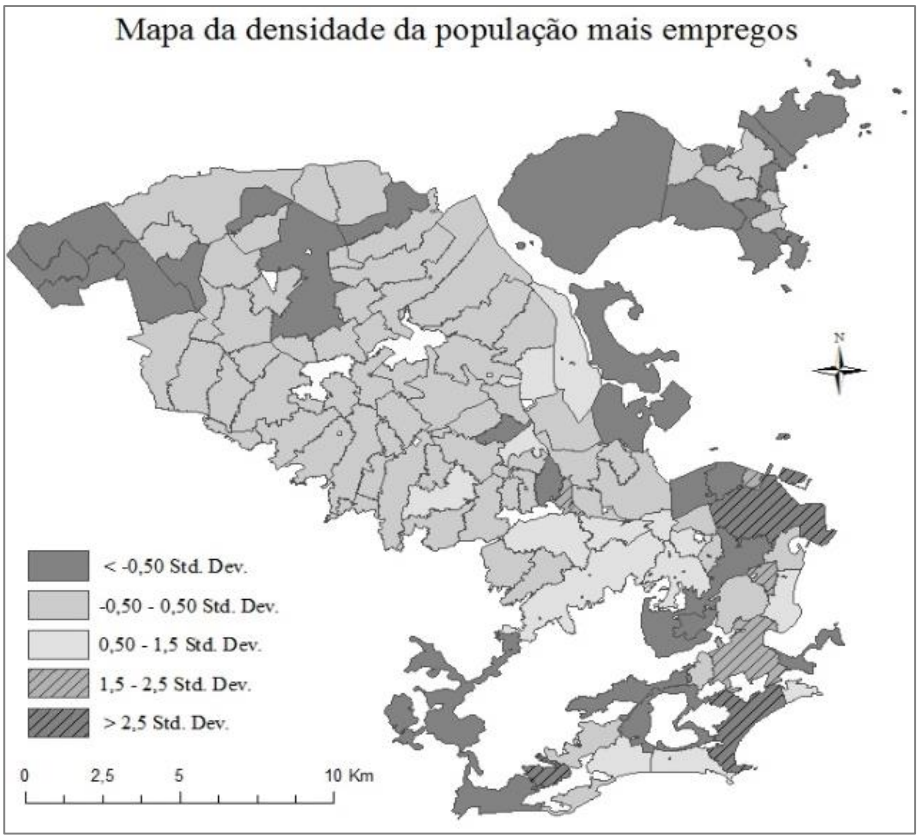


Figura 39 Mapas da variável densidade da população mais empregos

Tabela 10 Sumário estatístico das variáveis empregadas na modelagem

| | Média | Desv. Pad. | Mínimo | Máximo |
|---|---|---|---|---|
| **Densidade de acidentes** | 185,92 | 160,86 | 12,20 | 797,01 |
| **Hierarquia ponderada** | 1,55 | 0,37 | 1,00 | 3,18 |
| **Idade média** | 36,6 | 3,84 | 28,01 | 45,34 |
| **Densidade da população mais empregos** | 207,22 | 138,62 | 18,66 | 1113,05 |

Na etapa de modelagem serão aproveitadas, dentre as variáveis selecionadas nesta fase, aquelas que apresentem o nível de significância dentro do previsto pelo modelo e que forneçam uma contribuição considerável ao modelo.

## 5.7 Calibração dos modelos estatísticos

De modo a selecionar os modelos estatísticos a serem adotados, deve-se primeiramente selecionar os modelos não espaciais, começando do mais simples para o mais complexo. A partir da análise de resíduos da regressão não espacial, será verificada a existência da dependência espacial e, caso haja, serão aplicados os modelos espaciais para corrigir tal dependência espacial. A seguir serão apresentadas as principais etapas a serem seguidas nos processamentos dos modelos estatísticos.

Os pressupostos da regressão múltipla devem ser testados inicialmente na variável dependente e por fim nos resíduos da regressão. Verificou-se a normalidade da variável dependente por meio dos testes de normalidade de Anderson-Darling e Shapiro-Wilks, obteve-se para ambos testes um valor de p-valor menor que 0,05, o que mostra que a distribuição desta variável não é normal e, portanto, deve-se aplicar uma transformação sobre a mesma. Após a aplicação da transformação de Box e Cox, verificou-se que em ambos os testes o p-valor ficou bem superior a 0,05, o que mostra que esta variável transformada passou a ter uma distribuição normal.

É importante visualizar a existência dos *outliers* globais, bem como dos *outliers* espaciais na variável resposta, de modo a evitar que estes valores exerçam um efeito considerável sobre as estimativas do modelo de regressão. Após a verificação do *boxplot* obtido a partir da variável resposta transformada, cujos *outliers* inferior e superior possuem intervalo interquartílico multiplicado por 1,5, é possível verificar no *boxplot* que deixam de haver *outliers* globais, assim como outliers espaciais observando o diagrama de dispersão de Moran (Figura 40).

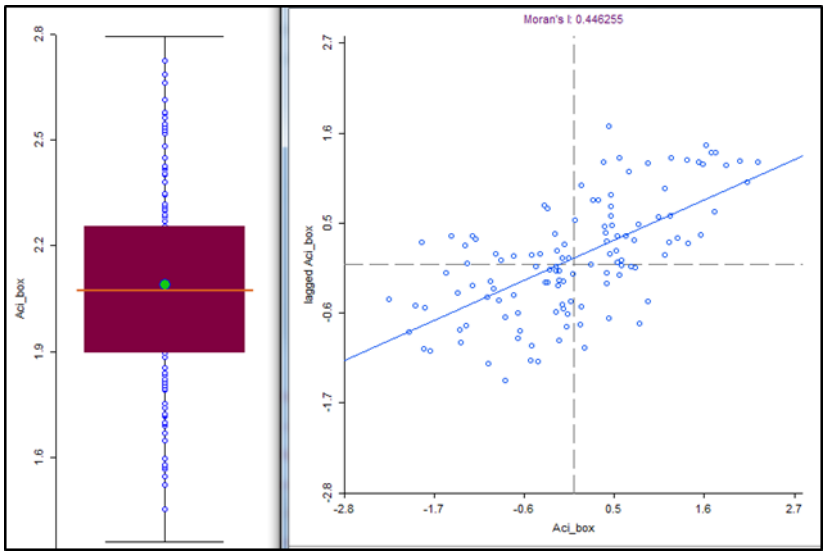


Figura 40 *Box plot* e diagrama de dispersão da variável resposta transformada

Como forma de verificar a relação entre a variável transformada (Acibox) e cada variável explicativa, geraram-se os gráficos constantes das Figuras 41 a 43.

É possível verificar que a hierarquia ponderada e a idade média contribuem individualmente ao modelo de regressão em mais de 40% enquanto que a densidade da população mais empregos contribui em pouco mais de 20%. No caso desta última variável, ao se ajustar uma linha de tendência logarítmica obteve-se um valor um pouco superior ao da linear (valor de $R^2$ em torno de 0,3). No entanto, quando inserida esta variável transformada pela função logarítmica no modelo estatístico, a mesma apresentada alta colinearidade com as outras duas variáveis.

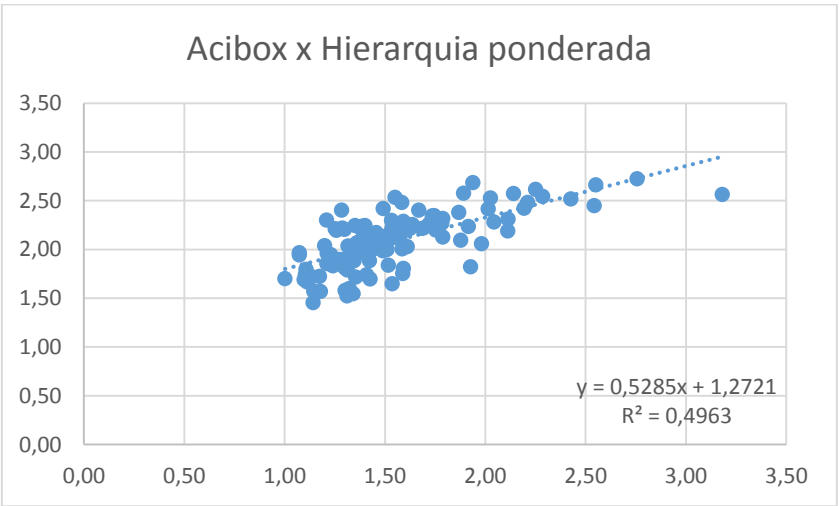


Figura 41 Gráfico da densidade de acidentes com transformação de Box e Cox versus hierarquia ponderada

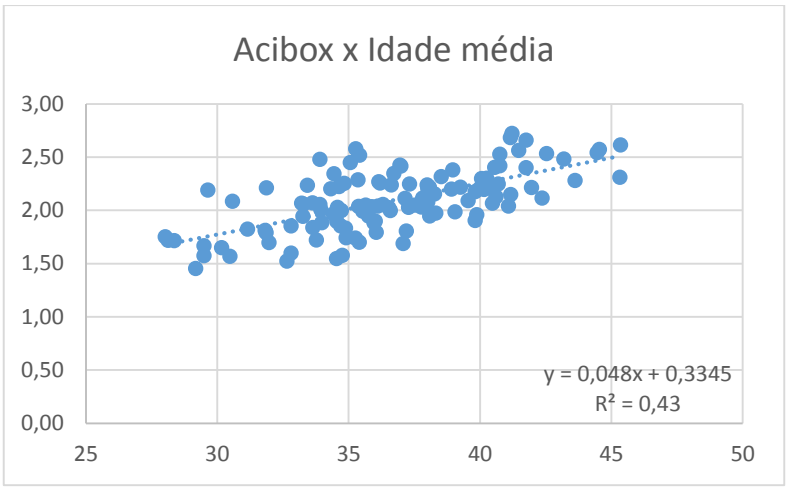


Figura 42 Gráfico da densidade de acidentes com transformação de Box e Cox versus idade média

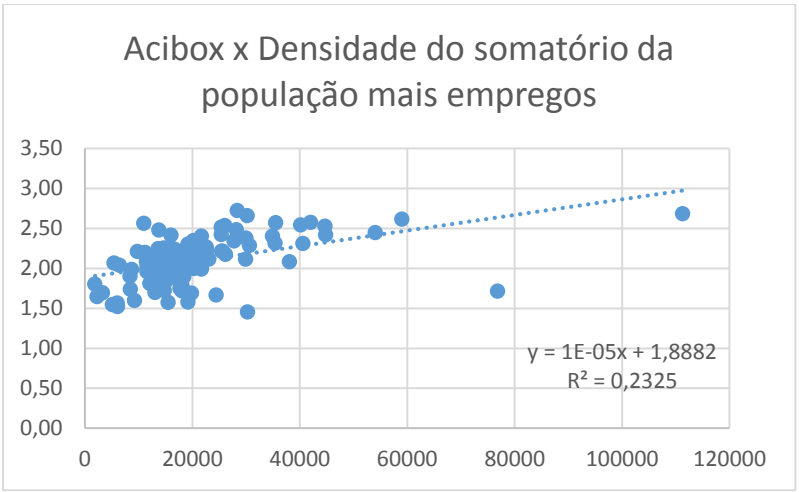


Figura 43 Gráfico da densidade de acidentes com transformação de Box e Cox versus somatório da população mais empregos

Após a transformação da variável resposta, processou-se o modelo de regressão múltipla com as variáveis explicativas selecionadas na etapa anterior, obtendo-se o sumário estatístico constante da Figura 44.

Observando o sumário estatístico é possível verificar que todas as variáveis explicativas apresentam p-valores significativos e que o valor do $R^2$ ajustado está em torno de 0,72 e o critério de informação de Akaike é de -110.

Outra grandeza a ser verificada na regressão é a multicolinearidade, a qual indica o nível de relação “perfeita” linear entre algumas ou todas as variáveis explicativas do modelo. A existência de multicolinearidade perfeita torna os coeficientes da regressão indeterminados e seus erros padrão infinitamente grandes. A multicolinearidade é detectada por meio de valores como é o caso do índice de condição de colinearidade, cuja interpretação empírica é que valores abaixo de 30 possuem colinearidade moderada a

forte e acima de 30, colinearidade grave. Observando o sumário, verifica-se a não existência de multicolinearidade grave entre as variáveis explicativas por se ter um índice de condição de multicolinearidade abaixo de 30. O valor do teste de Jarque-Bera não significativo mostra que os erros apresentam distribuição normal.

```
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :bairros_setores_271114.dbf
Weights matrix      :File: viz4_060115.gwt
Dependent Variable  :     Aci_box                Number of Observations:         119
Mean dependent var  :      2.0905                Number of Variables   :           4
S.D. dependent var  :      0.2819                Degrees of Freedom    :         115
R-squared           :      0.7262
Adjusted R-squared  :      0.7191
Sum squared residual:       2.568                F-statistic           :    101.6728
Sigma-square        :       0.022                Prob(F-statistic)     :   3.305e-32
S.E. of regression  :       0.147                Log likelihood        :      59.384
Sigma-square ML     :       0.022                Akaike info criterion :    -110.769
S.E of regression ML:      0.1469                Schwarz criterion     :     -99.652

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       0.2736851       0.1304047       2.0987362       0.0380286
           Hierpondlo       0.3570206       0.0401792       8.8857015       0.0000000
           Idade_med       0.0320703       0.0037788       8.4868066       0.0000000
            PopEmpha       0.0004346       0.0001044       4.1610869       0.0000614
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER           26.691

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2            1.700          0.4274

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                3            1.484          0.6859
Koenker-Bassett test              3            1.825          0.6094

SPECIFICATION ROBUST TEST
TEST                             DF        VALUE           PROB
White                             9            7.664          0.5683

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Moran's I (error)             0.1934           3.534          0.0004
Lagrange Multiplier (lag)          1           5.689          0.0171
Robust LM (lag)                    1           0.325          0.5684
Lagrange Multiplier (error)        1           9.970          0.0016
Robust LM (error)                  1           4.607          0.0318
Lagrange Multiplier (SARMA)        2          10.296          0.0058
```

Figura 44 Sumário estatístico da regressão múltipla

A Figura 45 mostra a análise de resíduo da regressão múltipla. Observando o gráfico Normal Q-Q, é possível observar que os valores dos resíduos apresentam valores próximos ao da distribuição normal. Pelos gráficos dos resíduos studentizados em função dos valores ajustados, é possível que não exista *outliers* na variável resposta e pela distância de Cook é possível verificar que não existem *outliers* entre as variáveis explicativas. Pelo gráfico dos resíduos em função dos resíduos ajustados é possível

verificar que não existe heterocedasticidade.

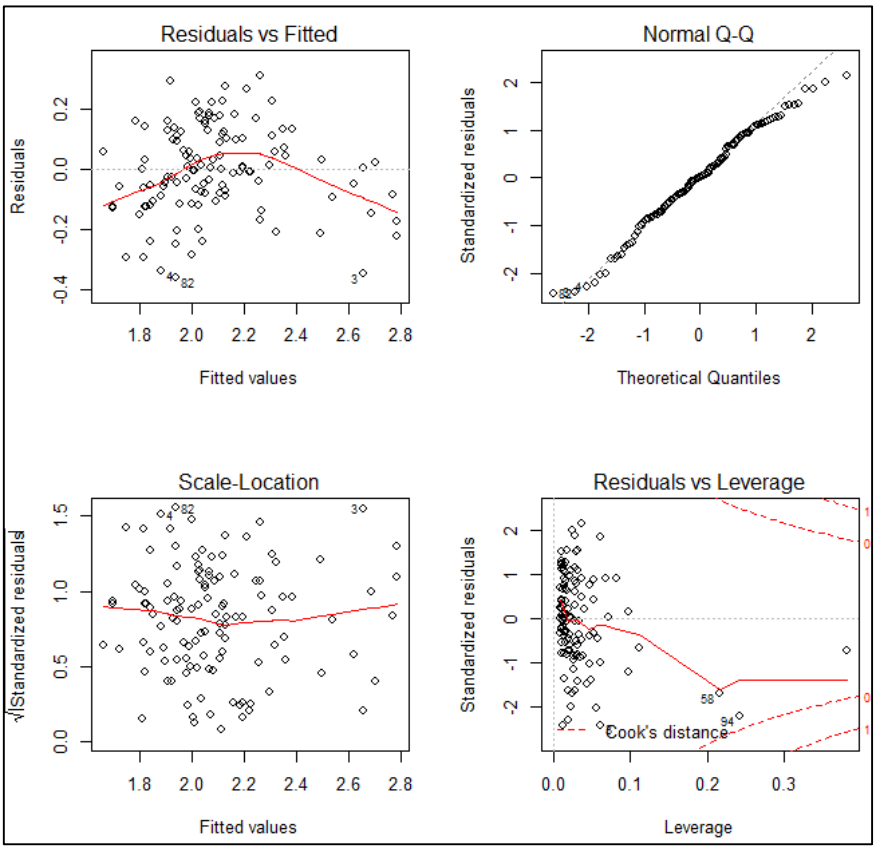


Figura 45 Gráficos dos resíduos da regressão múltipla

O mapa com os resíduos consta da Figura 46. No mapa constam em hachuras os bairros com os resíduos mais superestimados e em cinza mais escuro aqueles mais subestimados em valores absolutos. É possível verificar que somente oito bairros entre os 119 apresentam resíduos com desvio-padrão acima de dois em valores absolutos. Observa-se também que existem mais bairros com valores abaixo da média que acima, ou seja, a maior parte dos valores de densidade de acidentes foi superestimado. Por outro lado, sabe-se que grande parte destes bairros com maiores valores absolutos de desvio-padrão dos resíduos apresentam também os maiores valores absolutos da densidade de acidentes.

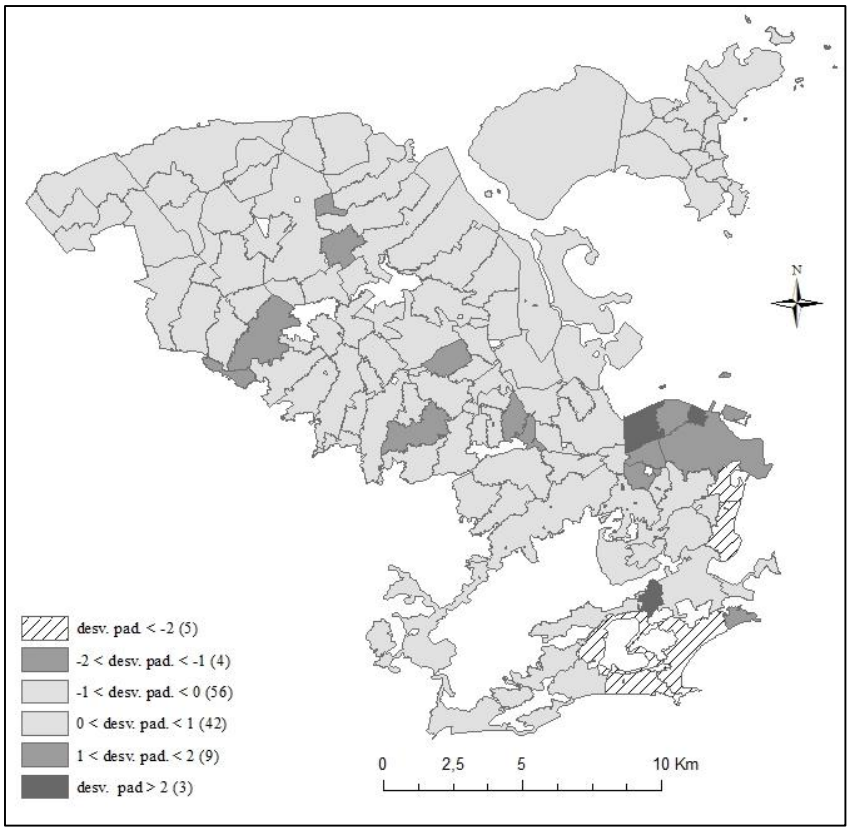


Figura 46 Mapa dos resíduos da regressão múltipla

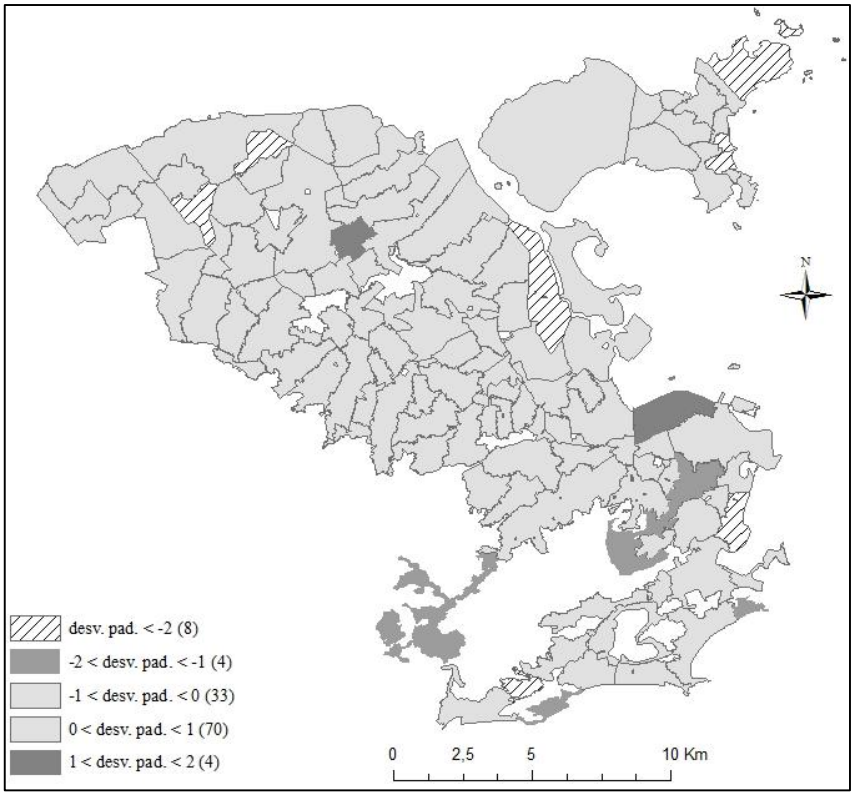


Figura 47 Mapa dos resíduos da regressão múltipla, ponderado pela densidade de acidentes

Como forma de analisar melhor os resíduos, dividiu-se o resíduo pela densidade de acidentes, obtendo-se o mapa constante da Figura 47. Observando-se este mapa, é possível perceber que ocorre uma alteração na distribuição espacial dos resíduos no mapa. Agora já constam com desvio-padrão maior que dois, em valores absolutos, os bairros com menores valores absolutos de densidade de acidentes. A única exceção neste caso foi o bairro do Flamengo, o qual consta como um bairro que possui grande densidade de acidentes mas que também possui uma previsão de acidentes com grande erro também.

Após os modelos de regressão múltipla, processou-se os MLG, começando por aqueles com distribuição de Poisson. No entanto, tendo em vista que este modelo é indicado para dados de contagem, empregou-se como variável resposta a média de acidentes nos anos de 2008 a 2010 e não a densidade de acidentes como ocorreu na regressão múltipla. Sabendo-se que estes não são adequados para o caso de haver superdispersão nos dados, aplicou-se antes um teste de superdispersão. Verificou-se a existência deste efeito, o que fez com que se empregasse a distribuição binomial negativa, cujo resultado encontra-se na Figura 48 e, os gráficos dos resíduos, na Figura 49.

```
Call:
glm.nb(formula = m ~ Idade_med + Hierpondlo + PopEmpha, data = fit1,
    init.theta = 1.775211533, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5917  -0.9709  -0.2111   0.3381   2.3360

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.0264774  0.6685640   1.535  0.12470
Idade_med   0.0802086  0.0193684   4.141 3.45e-05 ***
Hierpondlo  0.9373046  0.2054385   4.562 5.06e-06 ***
PopEmpha    0.0014248  0.0005345   2.666  0.00768 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.7752) family taken to be 1)

    Null deviance: 235.66  on 118  degrees of freedom
Residual deviance: 129.76  on 115  degrees of freedom
AIC: 1586.6
```

Figura 48 Resultado do modelo do MLG com distribuição binomial negativa

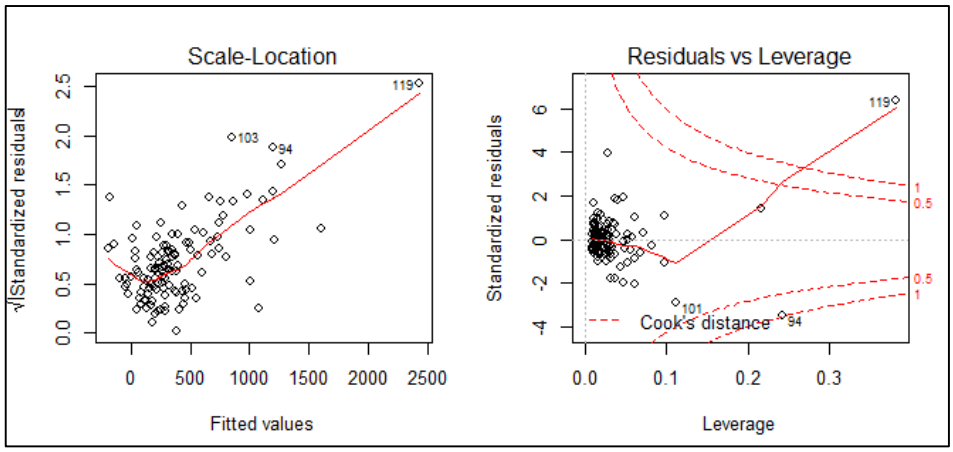


Figura 49 Gráficos dos resíduos do MLG com distribuição binomial negativa

É possível verificar no sumário que as variáveis independentes empregadas na regressão múltipla mostraram-se significativas também neste modelo. O valor do Akaike de 1586,6 é bem maior que o valor encontrado na regressão múltipla, embora não se possa empregar o mesmo Akaike para comparar modelos cuja variável resposta é diferente. Como forma de comparar os resultados do modelo de regressão com o MLG, calculou-se o erro médio quadrático destes modelos, obtendo-se os valores de 100 e 370, respectivamente, o que mostra um resultado bem melhor do primeiro modelo.

Observando-se os gráficos dos resíduos studentizados em função dos valores ajustados, é possível ver a existência de *outliers* na variável resposta presentes pelos bairros da Rocinha (código 94), Centro (código 119) e Tijuca (código 103). Ao se observar o gráfico da distância de Cook, é possível verificar *outliers* nas variáveis explicativas dos bairros Centro e Rocinha.

Processou-se os MLG com distribuição binomial negativa retirando-se alguns *outliers*. Em um primeiro momento, retirou-se somente o Centro e, em um segundo momento, retirou-se além do Centro também a Tijuca. Nestes modelos, a variável independente densidade da população mais emprego deixou de ser significativa. Os valores do erro médio quadrático obtidos foram de 343 e 290, respectivamente, o que mostra que houve uma certa melhora dos resultados do MLG em relação aos mesmos com os *outliers*, porém manteve-se em um patamar ainda bem pior que o da regressão múltipla.

Após a interpretação dos resultados do valor do ajuste do modelo de regressão e da significância das variáveis explicativas, deve-se observar a dependência espacial dos resíduos deste modelo. A Figura 44, que apresentou o sumário da regressão múltipla, possui um diagnóstico dos modelos da dependência espacial. No programa GeoDa Space,

deve-se selecionar uma matriz de proximidade para que seja calculado o diagnóstico da dependência espacial. É possível verificar que todas as variáveis explicativas são bastante significativas. É possível ver que o índice de Moran obtido foi de 0,19 e que, pela significância dos multiplicadores de Lagrange, os melhores modelos sugeridos por esses testes são os modelos os modelos CAR e SARMA, tendo em vista que apresentam significância no teste padrão e no robusto. No entanto, a definição do melhor modelo será dada somente após a verificação do valor dos resíduos e do índice de Moran. Aquele que tiver os menores valores de ambos os critérios deve ser escolhido.

Geraram-se então, neste programa, os modelos espaciais SAR, CAR e SARMA. Na pesquisa serão aplicados somente os dois primeiros modelos. Calculou-se o índice de Moran de todos os resíduos, obtendo-se valores de 0,20 e 0,28 respectivamente, o que mostra que estes modelos não diminuíram a dependência espacial. Quanto ao erro médio quadrático, os modelos SAR e CAR apresentaram os valores de 95 e 96, respectivamente. Este resultado mostra que os resultados dos modelos espaciais não são consideravelmente superiores ao modelo não espacial, cujo valor do erro médio quadrático foi de 100.

Como forma de verificar a contribuição de cada uma das variáveis explicativas, construiu-se um modelo de regressão simples para cada uma das variáveis e observou-se o ajuste do modelo e os efeitos espaciais dos resíduos. A Tabela 11 apresenta o resultado dos modelos de regressão simples com cada uma das variáveis explicativas.

Tabela 11 Resultados da regressão simples com cada uma das variáveis explicativas

| **Variáveis** | **$R^2$** | **Índice de Moran** | **Teste de White** |
|---|---|---|---|
| Hierarquia ponderada | 0,49 | 0,11 | 0,50* |
| Idade média | 0,43 | 0,47 | 0,27* |
| Densidade população mais emprego | 0,23 | 0,007 | 0,31* |
| Regressão com todas as variáveis | 0,72 | 0,19 | 0,57* |

(*) Valor da significância do teste.

Conforme se pode observar pelo valor do $R^2$, a hierarquia ponderada foi a variável que melhor se ajustou aos dados da densidade de acidentes. O resíduo da regressão simples com a idade média foi o que apresentou o maior valor do índice de Moran,

indicando maior dependência espacial. A densidade de população mais emprego foi a que apresentou menor valor do índice de Moran Comparando-se com os resultados da regressão com todas as variáveis, é possível verificar a influência que a hierarquia ponderada e a idade média exercem sobre o ajuste da regressão, assim como a idade média e a densidade da população maior o emprego exercem sobre a dependência espacial. Quanto à heterogeneidade espacial, aumenta com o aumento do valor da significância. Neste caso, a variável hierarquia ponderada foi a que apresentou maior homogeneidade dentre as variáveis explicativas.

Por fim, processaram-se os dados pelos regimes espaciais, conforme apresentado no item seguinte.

### 5.8 Análise da heterogeneidade espacial

Após a correção da dependência espacial (caso esteja presente), verifica-se a existência ou não da heterogeneidade espacial. Dentre os modelos que buscam modelar a heterogeneidade espacial, o adotado nesta pesquisa foi o de regimes espaciais, o qual divide a região de estudo de forma discreta, atribuindo valores diferentes aos coeficientes de cada regime. No contexto desta pesquisa, empregou-se o programa REDCAP de forma a dividir as regiões de trabalho. Quanto ao número de divisões, ou seja, de possíveis regimes, serão testadas as divisões em dois e três regiões. Não serão testados um número de divisões maior que quatro regiões pois daria um número pequeno de bairros por regime (abaixo de 30 bairros por regime em média, considerando um total de 119 bairros). É comum que nestes casos o programa escolha pelo menos um dos regimes com uma quantidade de bairros bem abaixo da média da quantidade de bairros por regime, o que aumentaria a variação dos dados entre os regimes, o que vai contra a ideia de homogeneidade buscada na pesquisa. Para evitar esse problema, quando se processou os dados para o caso de três regimes, colocou-se como número mínimo de regimes o de 25 para que não gerasse *clusters* com poucos bairros. Quando se tentou colocar 30 bairros como valor mínimo, verificou-se que o programa tendia a dividir a região em somente dois clusters e não em três para alguns métodos. Os métodos de regionalização testados estão apresentados na Tabela 12. Ao processar os métodos com dois clusters, verificou-se que o ALK tinha mesmo resultado que o CLK e que o SLK com vizinho mais próximo tinha o mesmo resultado que o SLK com a ligação completa.

Tabela 12 Métodos de regionalização empregados

| Número de divisões | Métodos |
|---|---|
| **2** | SLK prim. ordem |
| | SLK todas ordens |
| | ALK todas ordens |
| | CLK todas ordens |
| **3** | SLK prim. ordem |
| | SLK todas ordens |
| | ALK todas ordens |
| | CLK todas ordens |

A Figura 50 apresenta os mapas resultantes da aplicação de cada método de regionalização com duas regiões pelos métodos ALK e SLK, ambos com todas as ordens. A Figura 51 apresenta os mapas com três regiões pelos métodos SLK primeira ordem e SLK, ALK e CLK com todas as ordens.

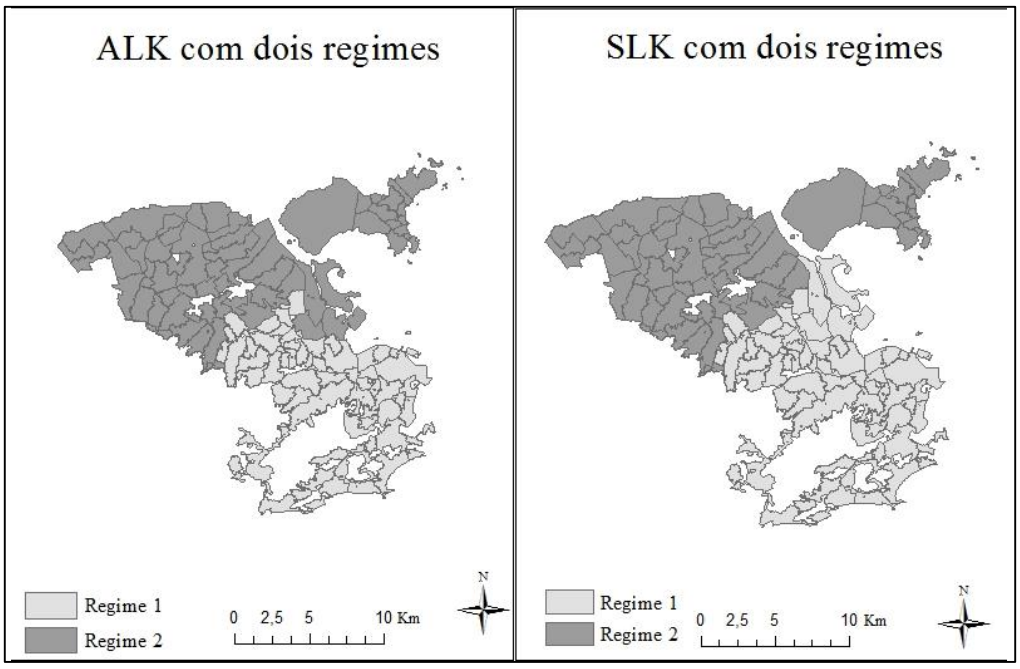


Figura 50 Mapas da região de trabalho dividida em duas regiões pelos métodos ALK e SLK com todas as ordens

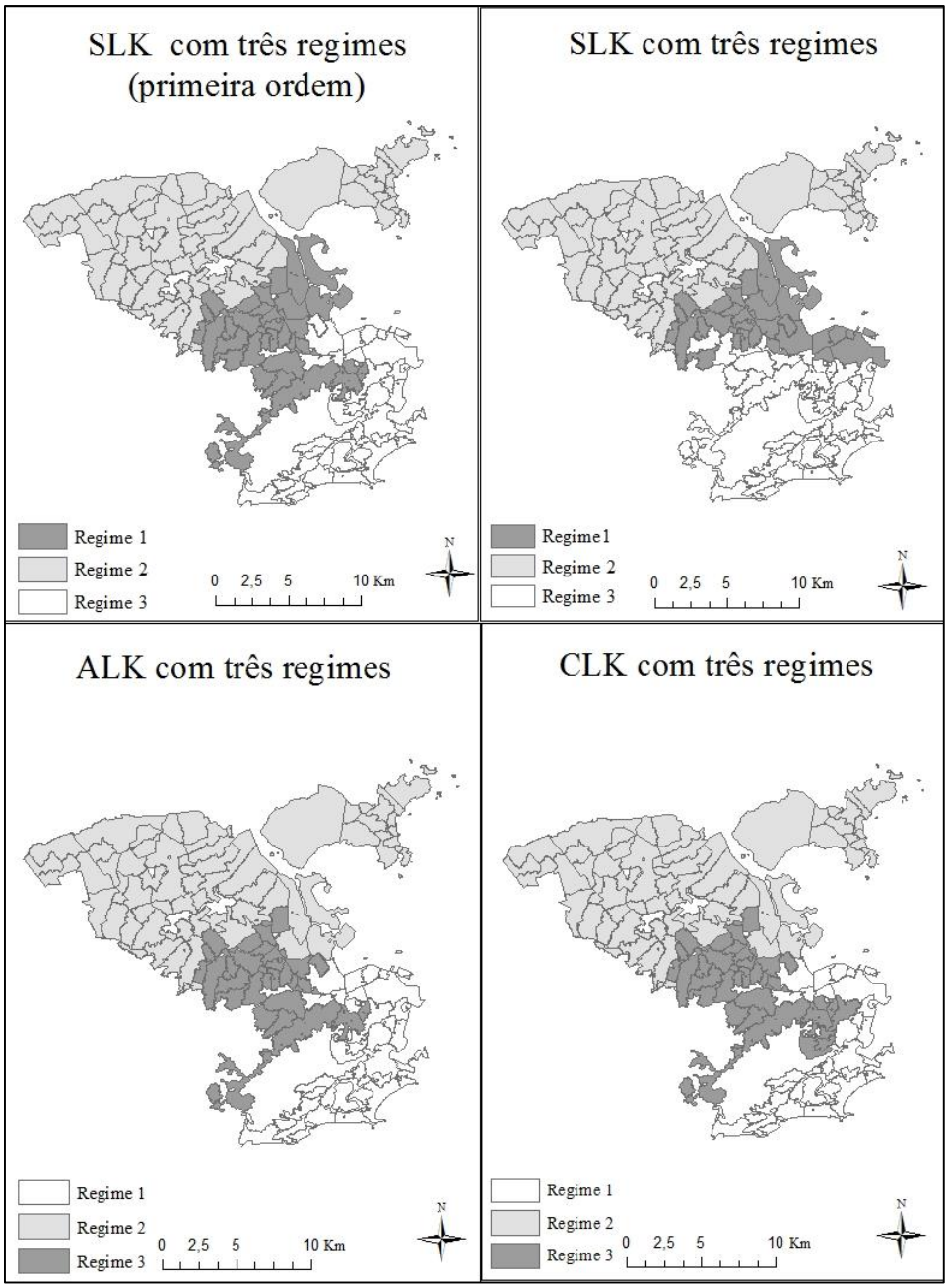


Figura 51 Mapas da região de trabalho dividida em três regiões pelos métodos SLK primeira ordem e SLK, ALK e CLK com todas as ordens

Retornando ao sumário do modelo de regressão na Figura 44, é possível observar que os testes de heterogeneidade apresentam significância bem acima de 0,05, o que sugere não haver heterogeneidade espacial. Mesmo assim, tendo em vista a análise exploratória sugerir a existência de um comportamento diferente da variável resposta em pelo menos dois locais, serão testados os modelos de regressão com regimes espaciais. Este processamento ocorrerá de três modos: considerando o campo da tabela referente ao regime (no qual consta valores 0 e 1 para caso de 2 regimes) como uma variável *dummy*,

processando os regimes conjuntamente e com os regimes em separado.

Ao se verificar os regimes com uma variável *dummy*, verificou-se que em nenhum dos casos houve melhora nos resultados do modelo de regressão que justificassem sua escolha. Como exemplo, a Figura 52 mostra o sumário do melhor resultado obtido, que foi o caso do método ALK com 2 regimes espaciais, cuja visualização espacial encontra-se na Figura 53. Neste caso o valor do $R^2$ajustado está em 0,73, bem próximo ao valor sem regimes espaciais e todas as variáveis explicativas são significativas.

```
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :bairros_setores_271114_redcap.dbf
Dependent Variable  :     Aci_box                Number of Observations:         119
Mean dependent var  :      2.0905                Number of Variables   :           5
S.D. dependent var  :      0.2819                Degrees of Freedom    :         114
R-squared           :      0.7356
Adjusted R-squared  :      0.7263
Sum squared residual:       2.480                F-statistic           :     79.2792
Sigma-square        :       0.021                Prob(F-statistic)     :   5.063e-32
S.E. of regression  :       0.144                Log likelihood        :      61.456
Sigma-square ML     :       0.021                Akaike info criterion :    -112.912
S.E of regression ML:      0.1444                Schwarz criterion     :     -99.016

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       0.4552948       0.1557138       2.9239210       0.0041705
          Hierpondlo       0.3124509       0.0450590       6.9342627       0.0000000
           Idade_med       0.0303109       0.0038112       7.9530892       0.0000000
            PopEmpha       0.0003833       0.0001056       3.6284184       0.0004282
              alk2_1      -0.0720841       0.0351070      -2.0532682       0.0423349
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER           31.357

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2           1.838          0.3990

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                4           1.232          0.8729
Koenker-Bassett test              4           1.422          0.8404

SPECIFICATION ROBUST TEST
Not computed due to multicollinearity.
```

Figura 52 Sumário do método ALK com 2 regimes como variável *dummy*

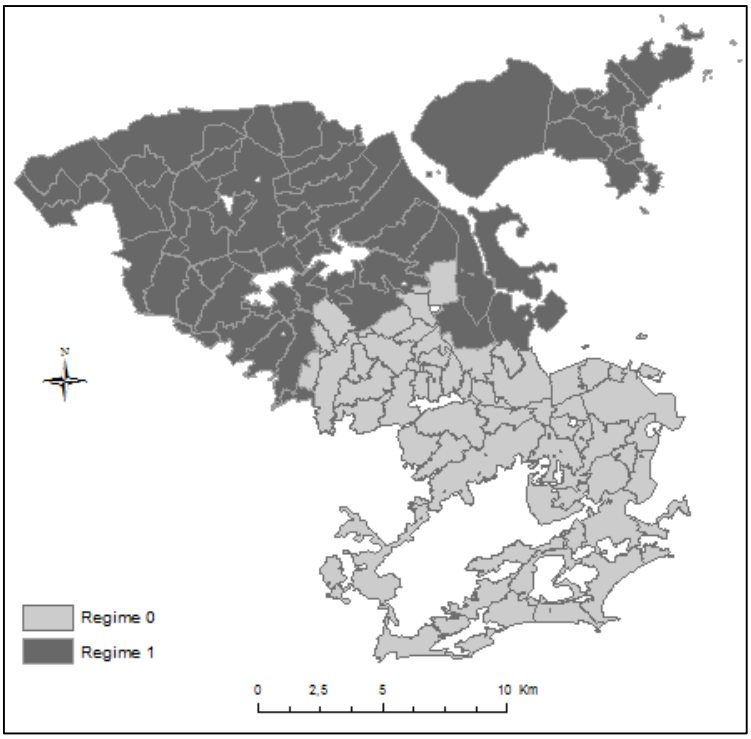


Figura 53 Regimes espaciais obtidos pelo método ALK

Ao se calcular os regimes espaciais, verificou-se que ao se processar conjuntamente, em todos os casos houve problema de colinearidade entre as variáveis. A Figura 54 mostra o exemplo do sumário para caso do método ALK com 2 regimes espaciais, onde o resultado do modelo obteve uma pequena do valor do $R^2$ajustado (0,76) em relação ao da regressão múltipla. No entanto, observando o número de condição de colinearidade acima de 30 (33, 267), é possível ver que as variáveis são consideradas colineares neste modelo.

```
Data set            :bairros_setores_271114_redcap.dbf
Weights matrix      :File: viz4_060115.gwt
Dependent Variable  :     Aci_box                Number of Observations:         119
Mean dependent var  :      2.0905                Number of Variables   :           8
S.D. dependent var  :      0.2819                Degrees of Freedom    :         111
R-squared           :      0.7710
Adjusted R-squared  :      0.7565
Sum squared residual:       2.148                F-statistic           :     53.3802
Sigma-square        :       0.018                Prob(F-statistic)     :   1.174e-32
S.E. of regression  :       0.134                Log likelihood        :      70.008
Sigma-square ML     :       0.018                Akaike info criterion :    -124.017
S.E of regression ML:      0.1344                Schwarz criterion     :    -101.784

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
          0_CONSTANT       0.8309536       0.1805410       4.6025751       0.0000112
        0_Hierpondlo       0.2981420       0.0459574       6.4873531       0.0000000
         0_Idade_med       0.0214525       0.0045677       4.6965811       0.0000076
          0_PopEmpha       0.0003386       0.0001033       3.2790457       0.0013915
          1_CONSTANT      -0.3694234       0.2285818      -1.6161545       0.1088996
        1_Hierpondlo       0.4439192       0.1036784       4.2816943       0.0000396
         1_Idade_med       0.0432065       0.0056504       7.6466386       0.0000000
          1_PopEmpha       0.0011897       0.0003366       3.5348667       0.0005962
------------------------------------------------------------------------------------
Regimes variable: alk2_1

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER           33.267

TEST ON NORMALITY OF ERRORS
TEST                             DF         VALUE           PROB
Jarque-Bera                       2           5.973          0.0505

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF         VALUE           PROB
Breusch-Pagan test                7          12.648          0.0812
Koenker-Bassett test              7          11.423          0.1212

SPECIFICATION ROBUST TEST
Not computed due to multicollinearity.

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Lagrange Multiplier (lag)         1           2.751          0.0972
Robust LM (lag)                   1           0.080          0.7768
Lagrange Multiplier (error)       1           8.201          0.0042
Robust LM (error)                 1           5.530          0.0187
Lagrange Multiplier (SARMA)       2           8.281          0.0159
```

Figura 54 Sumário do método ALK com 2 regimes espaciais

No caso deste modelo, ao se observar a significância do teste de Chow (Figura 55), verifica-se que a hierarquia ponderada apresenta valor de 0,1986, sendo considerada não significante para um nível de confiança de 95%, ou seja, é a única variável que possui o mesmo comportamento em ambos os regimes espaciais.

```
REGIMES DIAGNOSTICS - CHOW TEST
                VARIABLE         DF        VALUE           PROB
                CONSTANT          1         16.983          0.0000
              Hierpondlo          1          1.652          0.1986
               Idade_med          1          8.964          0.0028
                PopEmpha          1          5.845          0.0156
             Global test          4         23.263          0.0001
```

Figura 55 Teste de Chow do método ALK com dois regimes espaciais

As Figuras 56 e 57 mostram os sumários dos regimes separadamente para os regimes

0 e 1, ou seja, região sul e norte, respectivamente. É possível perceber que o ajuste de cada região separadamente é inferior ao ajuste do processamento em conjunto e da regressão múltipla sem os regimes espaciais. É possível verificar também que o ajuste na região sul ($R^2$ ajustado= 0,59)é inferior ao do regime mais ao norte ($R^2$ = 0,63).No entanto, na região sul o número de condição de colinearidade é de 26,553, enquanto que na região norte é de 33,267, o que faz com que na região norte as variáveis apresentem multicolinearidade grave, o que não ocorre na região sul.

```
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION - REGIME 0
-----------------------------------------------------------------
Data set            :bairros_setores_271114_redcap.dbf
Weights matrix      :File: viz4_060115.gwt
Dependent Variable  :   0_Aci_box               Number of Observations:          57
Mean dependent var  :      2.2714               Number of Variables   :           4
S.D. dependent var  :      0.2290               Degrees of Freedom    :          53
R-squared           :      0.6113
Adjusted R-squared  :      0.5892
Sum squared residual:       1.142               F-statistic           :     27.7784
Sigma-square        :       0.020               Prob(F-statistic)     :   6.229e-11
S.E. of regression  :       0.142               Log likelihood        :      30.560
Sigma-square ML     :       0.020               Akaike info criterion :     -53.120
S.E of regression ML:      0.1416               Schwarz criterion     :     -44.948

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
          0_CONSTANT       0.8309536       0.1902088       4.3686403       0.0000585
        0_Hierpondlo       0.2981420       0.0484184       6.1576209       0.0000001
         0_Idade_med       0.0214525       0.0048123       4.4578683       0.0000433
         0_PopEmpha        0.0003386       0.0001088       3.1123818       0.0029888
------------------------------------------------------------------------------------
Regimes variable: alk2

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER           26.553

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2           4.277          0.1179

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                3           6.321          0.0970
Koenker-Bassett test              3           5.758          0.1240

SPECIFICATION ROBUST TEST
TEST                             DF        VALUE           PROB
White                             9          21.537          0.0105
```

Figura 56 Sumário do método ALK na região denominada regime 0

```
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION - REGIME 1
-----------------------------------------------------------
Data set            :bairros_setores_271114_redcap.dbf
Weights matrix      :File: viz4_060115.gwt
Dependent Variable  :   1_Aci_box                Number of Observations:          62
Mean dependent var  :      1.9243                Number of Variables   :           4
S.D. dependent var  :      0.2167                Degrees of Freedom    :          58
R-squared           :      0.6488
Adjusted R-squared  :      0.6307
Sum squared residual:       1.006                F-statistic           :     35.7207
Sigma-square        :       0.016                Prob(F-statistic)     :   3.305e-13
S.E. of regression  :       0.127                Log likelihood        :      39.779
Sigma-square ML     :       0.016                Akaike info criterion :     -71.558
S.E of regression ML:      0.1274                Schwarz criterion     :     -63.049

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
          1_CONSTANT      -0.3694234       0.2167195      -1.7046152       0.0936180
        1_Hierpondlo       0.4439192       0.0982980       4.5160543       0.0000314
         1_Idade_med       0.0432065       0.0053572       8.0651799       0.0000000
          1_PopEmpha       0.0011897       0.0003191       3.7283488       0.0004396
------------------------------------------------------------------------------------
Regimes variable: alk2
Warning: The regimes operation resulted in islands for regime 1.

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER           33.267

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2           1.514           0.4692

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                3           5.197           0.1579
Koenker-Bassett test              3           4.821           0.1854

SPECIFICATION ROBUST TEST
Not computed due to multicollinearity.
```

Figura 57 Sumário do método ALK na região denominada regime 1

## 5.9 MAUP – geração de dados

A geração dos dados com vistas a verificar o MAUP é feita dividindo a área de trabalho em 30, 40, 50, 60, 70, 80, 90, 100, 110 e 119 regiões, sendo que este último corresponde aos dados originais. Julgou-se que um intervalo menor que 10 entre as regiões produziria resultados muito próximos entre si gerando informação redundante. Para que não se gere uma quantidade excessiva de dados, escolheu-se somente um método de regionalização, que foi o SLK com todas as ordens. As Figuras 58 a 60 apresentam os mapas resultantes das diversas divisões feitas na área de trabalho. Cada um destes mapas possui uma tabela de atributos com os valores das variáveis explicativas calculadas a partir dos dados originais dos 119 bairros por meio da operação espacial de dissolução de polígonos.

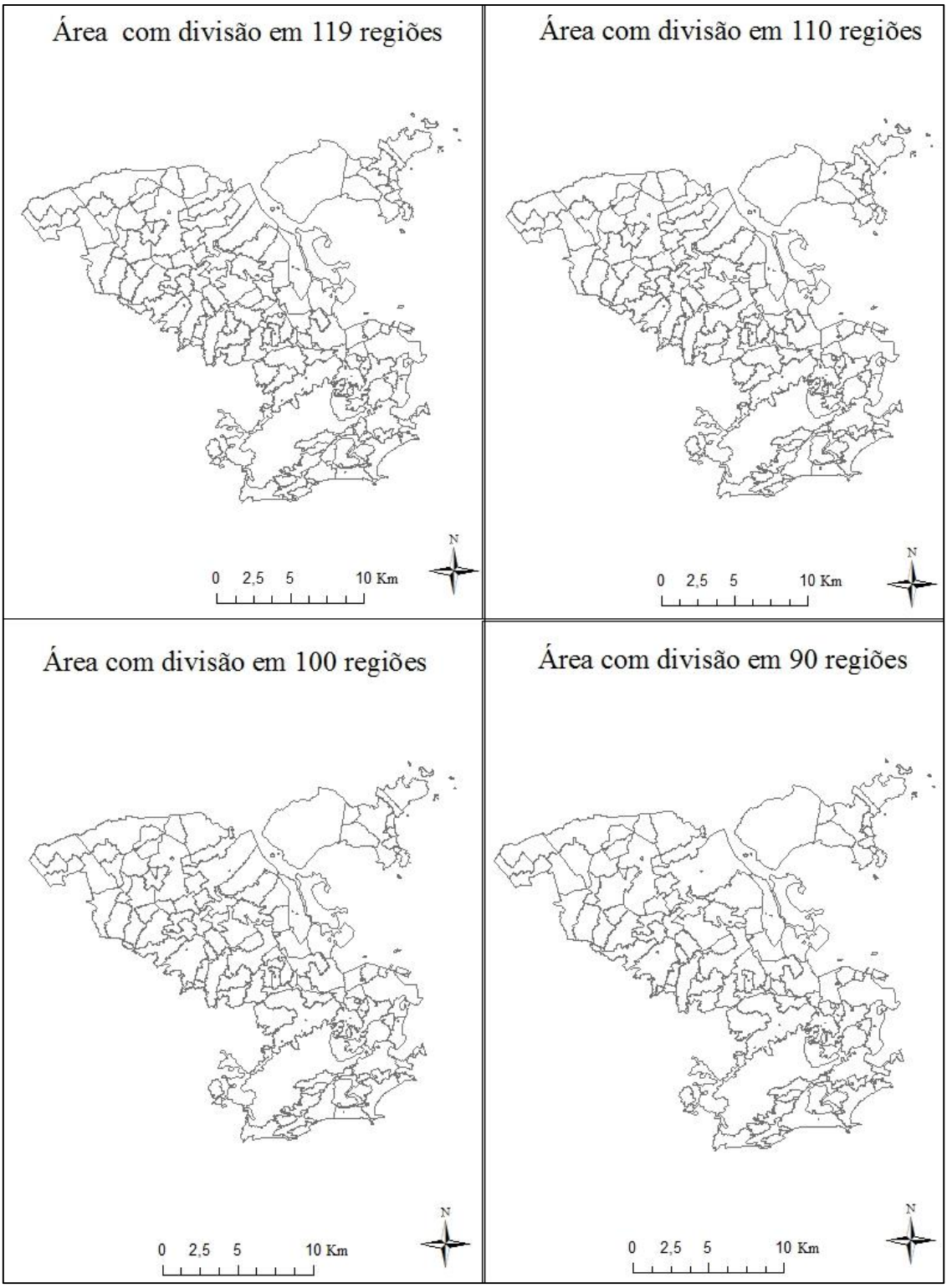


Figura 58 Mapas com a divisão da área de trabalho em 118, 110, 100 e 90 regiões

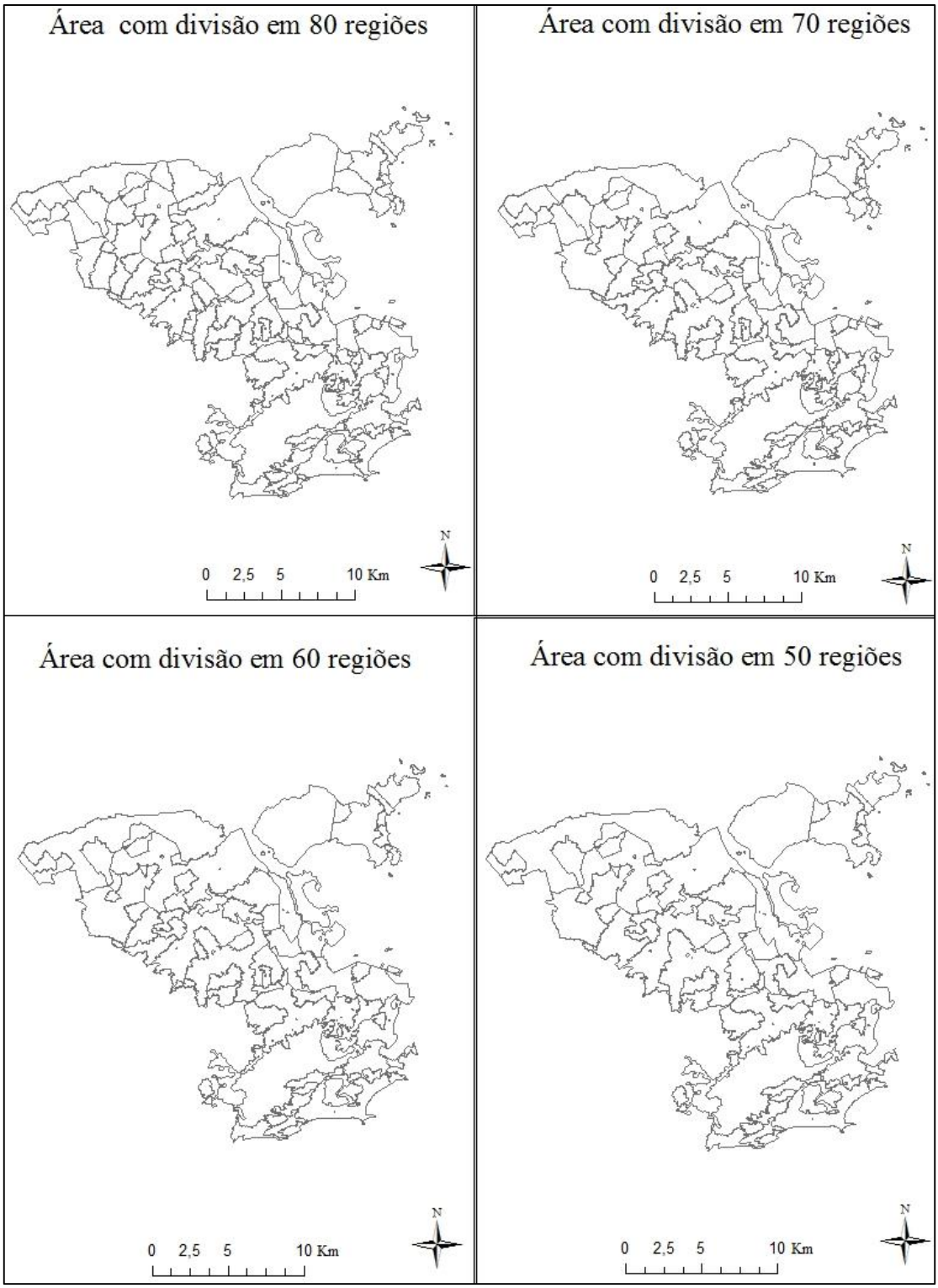


Figura 59 Mapas com a divisão da área de trabalho em 80, 70, 60 e 50 regiões

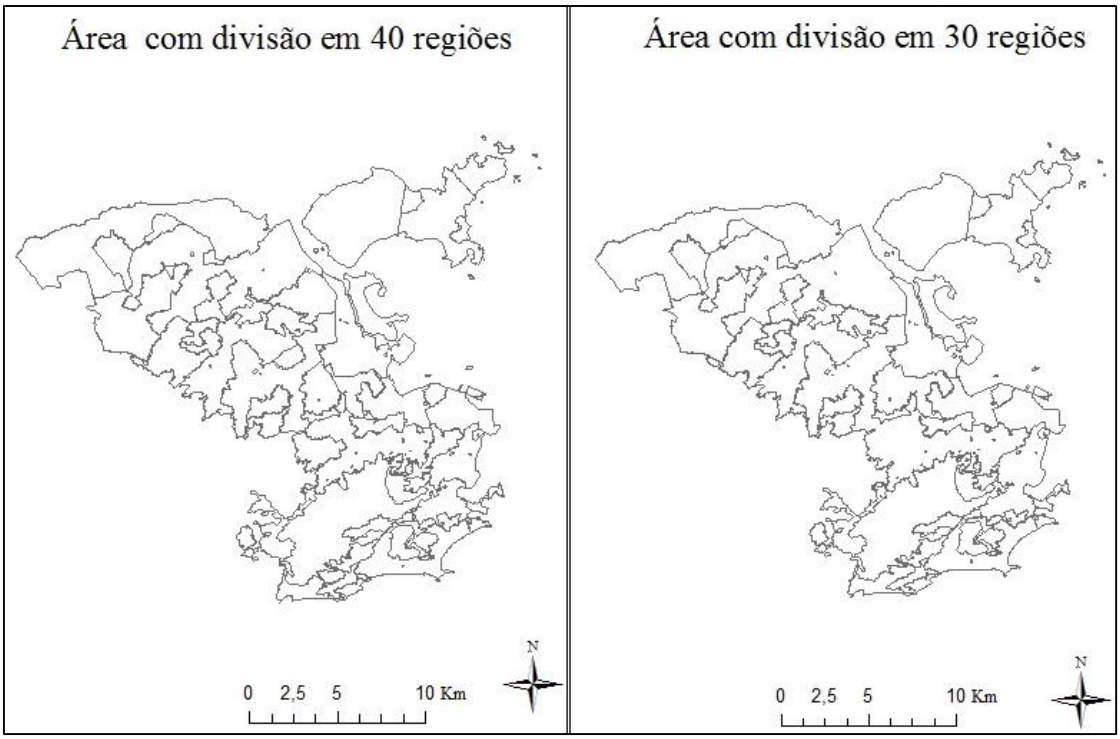


Figura 60 Mapas com a divisão da área de trabalho em 40 e 30 regiões

## 5.10 MAUP – Análise de sensibilidade

Após a geração dos diferentes mapas mostrados no item anterior, iniciou-se a etapa de modelagem dos dados com as diversas agregações. Ao processar o modelo de regressão múltipla para cada uma das agregações, verificou-se que em nenhuma delas a variável dependente apresentava distribuição normal, o que demandou a transformação de Box e Cox em todos os casos. Os valores de λ obtidos de em cada uma das agregações está mostrado da Tabela 13.

Tabela 13 Valores de λ obtidos de em cada um dos níveis de agregação

| **Agregações** | **λ** | **Agregações** | **λ** |
|---|---|---|---|
| **30** | 0,08 | **80** | 0,08 |
| **40** | 0,11 | **90** | 0,12 |
| **50** | 0,07 | **100** | 0,12 |
| **60** | 0,10 | **110** | 0,13 |
| **70** | 0,11 | **119** | 0,14 |

Como se pode observar os valores de λ são muito próximos. Por essa razão decidiu-se tomar o valor médio de 0,11 e aplicar tal transformação em todos os casos. Os

coeficientes da modelagem encontram na Tabela 14 e os gráficos de dispersão dos coeficientes das variáveis explicativas na Figura 61. Nestes gráficos, os pontos representam o valor da média do coeficiente e a barra vertical o erro padrão do mesmo.

Tabela 14 Valores da média e do erro padrão dos coeficientes das variáveis explicativas obtidos dos modelos de regressão

| | **Hierarquia ponderada** | | **Idade média** | | **Dens. Pop e Emp** | |
|---|---|---|---|---|---|---|
| **Agreg** | **Média** | **Erro pad.** | **Média** | **Erro pad.** | **Média** | **Erro pad.** |
| **30** | 0,240536 | 0,067201 | 0,020298 | 0,005732 | 0,028409 | 0,012564 |
| **40** | 0,246042 | 0,053739 | 0,024238 | 0,004278 | 0,020700 | 0,009003 |
| **50** | 0,227571 | 0,04753 | 0,024832 | 0,003529 | 0,022795 | 0,008422 |
| **60** | 0,229795 | 0,043932 | 0,024703 | 0,003234 | 0,022963 | 0,007949 |
| **70** | 0,251439 | 0,04092 | 0,021962 | 0,003179 | 0,023388 | 0,008082 |
| **80** | 0,240967 | 0,032601 | 0,02106 | 0,002828 | 0,021389 | 0,007423 |
| **90** | 0,242069 | 0,03114 | 0,020991 | 0,002694 | 0,021720 | 0,007156 |
| **100** | 0,239815 | 0,029096 | 0,019478 | 0,002564 | 0,021436 | 0,007038 |
| **110** | 0,235794 | 0,02759 | 0,019649 | 0,002451 | 0,021669 | 0,006836 |
| **119** | 0,210041 | 0,025013 | 0,019619 | 0,002353 | 0,025515 | 0,006503 |

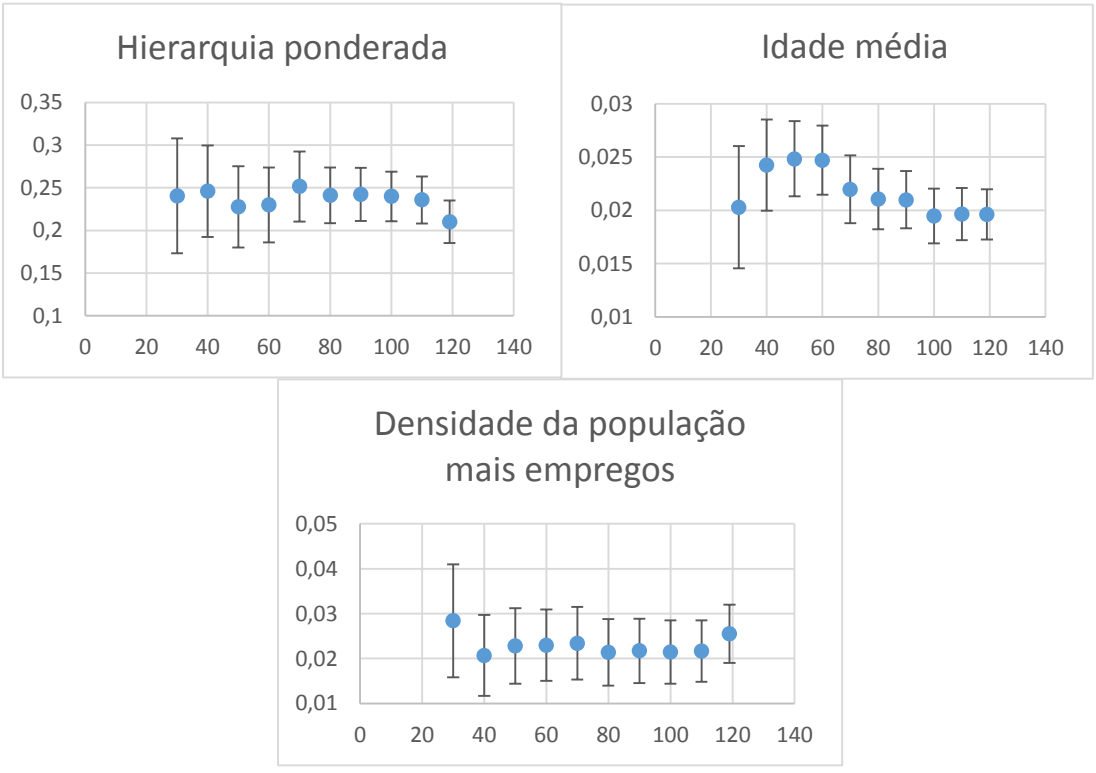


Figura 61 Gráficos da média e do erro padrão dos coeficientes das variáveis explicativas

Observando os gráficos, é possível verificar a ocorrência de maior estabilidade nas médias dos intervalos entre os níveis de agregação 70 e 110. O nível de agregação de 30 apresenta a menor estabilidade tanto na média do coeficiente quanto no valor do erro padrão. No entanto, é preciso aplicar um teste de hipóteses de modo a verificar se as diferenças entre as médias podem ser consideradas nulas ou não. Primeiramente, aplicou-se um teste para variância com a distribuição de Fisher-Snedecor a cada duas áreas de agregação adjacentes, de modo a verificar se as variâncias seriam iguais ou não. Como resultado, verificou-se que todas as variâncias eram consideradas iguais a um nível de confiança de 95%. Em seguida, aplicou-se o teste de hipótese para as diferenças das médias cujas variâncias são conhecidas e iguais, obtendo-se coeficientes considerados diferentes a um nível de confiança de 95% somente na variável idade média entre as agregações 30 e 40 e as agregações 60 e 70.

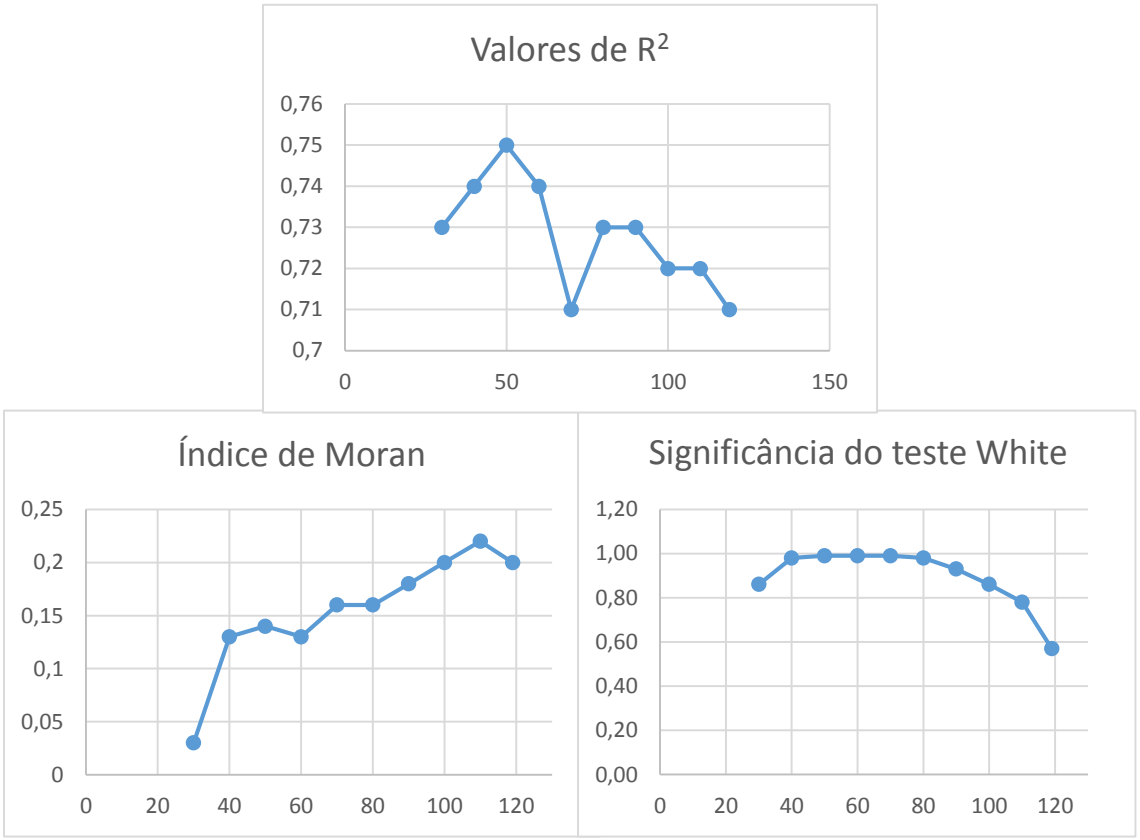


Figura 62 Gráficos dos valores de $R^2$, índice de Moran e teste de White dos resíduos da regressão

Além de se observar os valores da média e erro padrão também é preciso verificar o nível de significância das variáveis (neste caso de 95%) em todos os modelos. Após a observação desse valor nos modelos, verificou-se que o nível de significância esteve sempre abaixo de 95% em todas as variáveis e em todos os modelos. Outras grandezas

observadas foram o valor do $R^2$, o índice de Moran e o teste de White para verificação da heterogeneidade espacial. Os gráficos destes testes para cada um dos níveis de agregação encontram-se na Figura 62. Como se pode observar, o valor do $R^2$ sofre uma certa oscilação nos diferentes níveis de agregação, o índice de Moran dos resíduos aumenta à medida que aumenta o número de regiões enquanto que o teste de White mostra uma significância decrescente, mas mesmo assim ainda muito superior a 0,05 adotado na pesquisa como indicador de heterocedasticidade.

### 5.11 Validação dos modelos

Conforme mencionado, adotou-se como variável dependente a média da densidade de acidentes nos anos de 2008 a 2010. A validação do modelo foi feita com os dados de acidentes no ano de 2011. A densidade dos acidentes em 2011 está apresentada na Figura 63. Sabe-se que existe uma variação entre o número de acidentes empregado na calibração e na validação e essa variação não é uniforme de um bairro para outro. A Figura 64 mostra o mapa com a variação na densidade de acidentes entre os bairros dos dois conjuntos de dados (cujo valor é igual ao da variação da quantidade de acidentes). Comparando-se os mapas das Figuras 63 e 64, percebe-se que apresentam maior desvio padrão absoluto no mapa de variação da densidade de acidentes aqueles bairros com menores valores de densidade de acidentes. Os bairros com maior quantidade de acidentes, por sua vez, obtiveram valores mais próximos da média.

As variáveis explicativas da validação serão as mesmas que as da modelagem, tendo em vista não haver alteração considerável nas variáveis explicativas. Para se ter uma ideia, a população da cidade do Rio de Janeiro vem crescendo a uma taxa de cerca de 0,6% ao ano e idade média aumentou no Brasil de 31,8 para 34,7 anos de idade entre os anos de 2000 e 2010 segundo os Censos de 2000 e 2010 do IBGE, o que neste caso fornece uma taxa abaixo de 10% na década. Da mesma forma, segundo o relatório sobre a evolução das características do emprego e da economia da cidade do Rio de Janeiro (SILVA *et al.*, 2011), o número de estabelecimentos obteve um aumento de cerca de 12% em toda a década e renda média 14,2% em média no mesmo período na cidade do Rio de Janeiro. A variável que apresentou um aumento mais acentuado no período de 2000 a 2009 foi o número de empregos formais, que foi de 28,8%.

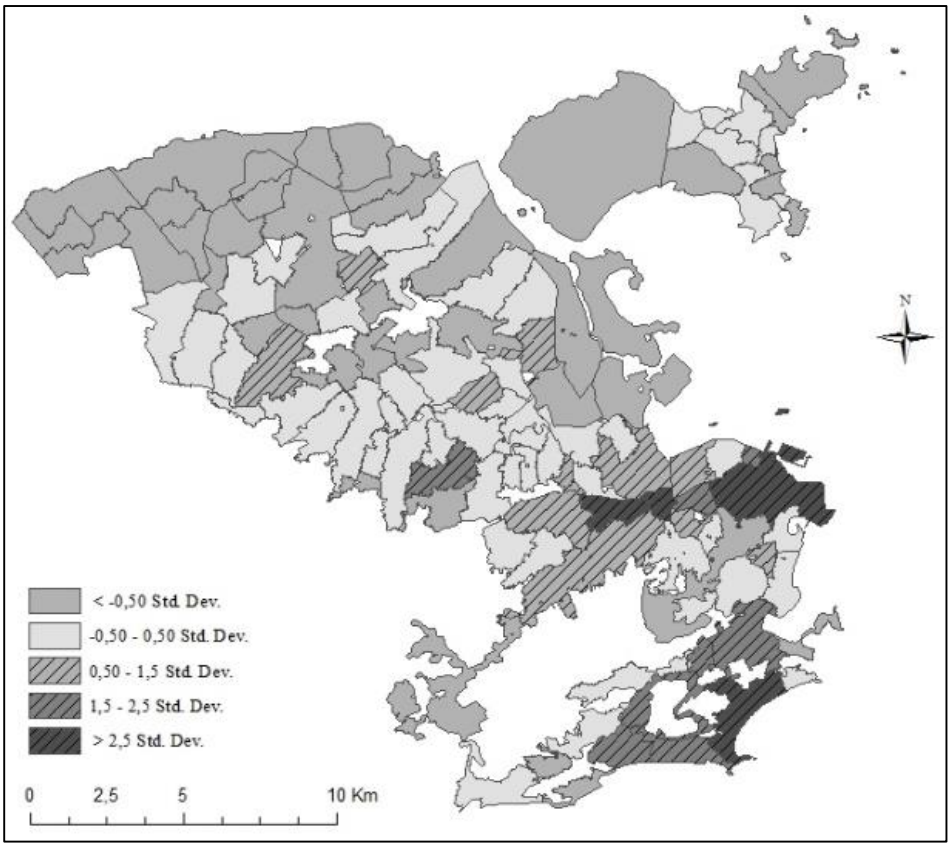


Figura 63 Mapa da densidade de acidentes da validação

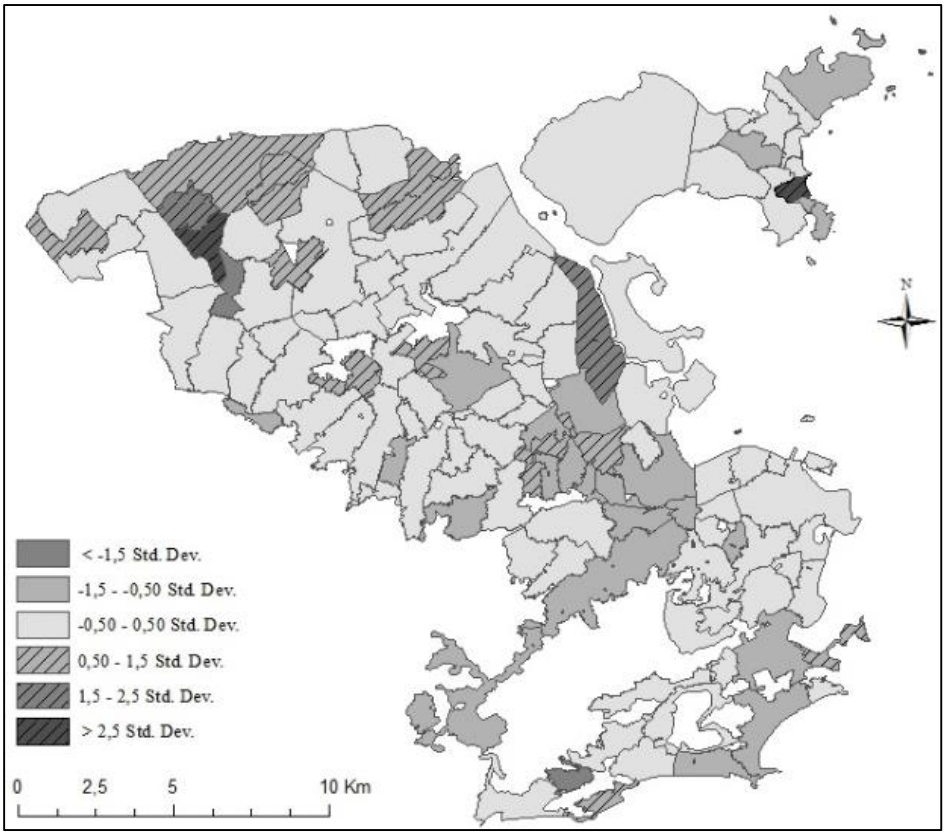


Figura 64 Mapas das variações da densidade de acidentes em relação aos valores da calibração

A Tabela 15 mostra o sumário estatístico da densidade de acidentes da calibração e a da validação. Como se pode observar, houve um aumento da ordem de 16% na média da densidade de acidentes entre os dados empregados na modelagem e os do ano de 2011, assim como nos valores do desvio padrão, valor mínimo e valor máximo.

Tabela 15 Sumário estatístico da densidade de acidentes da validação e da densidade empregada na modelagem

| | **Média** | **Desv. Pad.** | **Min** | **Max** |
|---|---|---|---|---|
| **Densidade de acidentes validação** | 216,46 | 176,62 | 19,77 | 818,10 |
| **Densidade de acidentes calibração** | 185,92 | 160,86 | 12,20 | 797,01 |

De forma a se ter uma ideia global do quão o valor previsto no modelo de validação se adequa aos valores observados, determinou-se o coeficiente de Pearson entre estas variáveis, obtendo-se o valor de 0,809. Comparando-se com o valor do Pearson entre a variável dependente do modelo de calibração e o valor estimado pelo modelo que é de 0,843, houve uma ligeira piora no valor mas mesmo assim valores bem aceitáveis. O valor do NRMS da validação ficou em 0,147 e o valor do RMS é de 117. O valor do NRMS obtido na calibração foi de 0,126 e o RMS de 99. Embora tenha sido inferior ao da calibração, o resultado da validação apresenta mesmo assim um bom valor de NRMS.

Os valores da média e erro padrão das variáveis obtidos na validação constam da Tabela 16. Mesmo que os valores dos coeficientes tenham sido considerados iguais quando da calibração, como forma de comparar os resultados da validação com os da calibração, geraram-se os gráficos das variáveis explicativas, os quais constam na Figura 65. Como se pode observar, os gráficos de cada uma das variáveis na calibração e na validação apresentam aspectos semelhantes. Da mesma forma que na calibração, os coeficientes da validação apresentaram valores estatisticamente diferentes entre os níveis de agregação 30 e 40, bem como entre as agregações 60 e 70. Os valores do $R^2$, índice de Moran e teste de White para os níveis de agregação utilizados, constam da Tabela 17 e os respectivos gráficos constam na Figura 66.

Tabela 16 Valores da média e do erro padrão dos coeficientes das variáveis explicativas obtidos dos modelos de regressão da validação

| | Hierarquia ponderada | | Idade média | | Dens. Pop e Emp | |
|---|---|---|---|---|---|---|
| **Agreg** | **Média** | **Erro pad.** | **Média** | **Erro pad.** | **Média** | **Erro pad.** |
| **30** | 0,2140303 | 0,0613331 | 0,0186982 | 0,0052249 | 0,0285718 | 0,0114447 |
| **40** | 0,2312626 | 0,0517329 | 0,0240535 | 0,0041169 | 0,0150455 | 0,0086577 |
| **50** | 0,2132155 | 0,0466749 | 0,0247993 | 0,0034655 | 0,0178019 | 0,008268 |
| **60** | 0,2236514 | 0,0444393 | 0,0239104 | 0,0032752 | 0,0193396 | 0,0080468 |
| **70** | 0,2501137 | 0,0407709 | 0,0209357 | 0,0031685 | 0,0199086 | 0,0080574 |
| **80** | 0,2286054 | 0,032624 | 0,0205724 | 0,00283 | 0,0182457 | 0,0074292 |
| **90** | 0,2292614 | 0,0312868 | 0,0203803 | 0,002706 | 0,0186162 | 0,0071893 |
| **100** | 0,2287764 | 0,029163 | 0,0186764 | 0,0025686 | 0,017862 | 0,0070391 |
| **110** | 0,227722 | 0,0275608 | 0,0184622 | 0,0024468 | 0,0179542 | 0,0068236 |
| **119** | 0,1933743 | 0,0259862 | 0,0196778 | 0,0024414 | 0,0292488 | 0,0067492 |

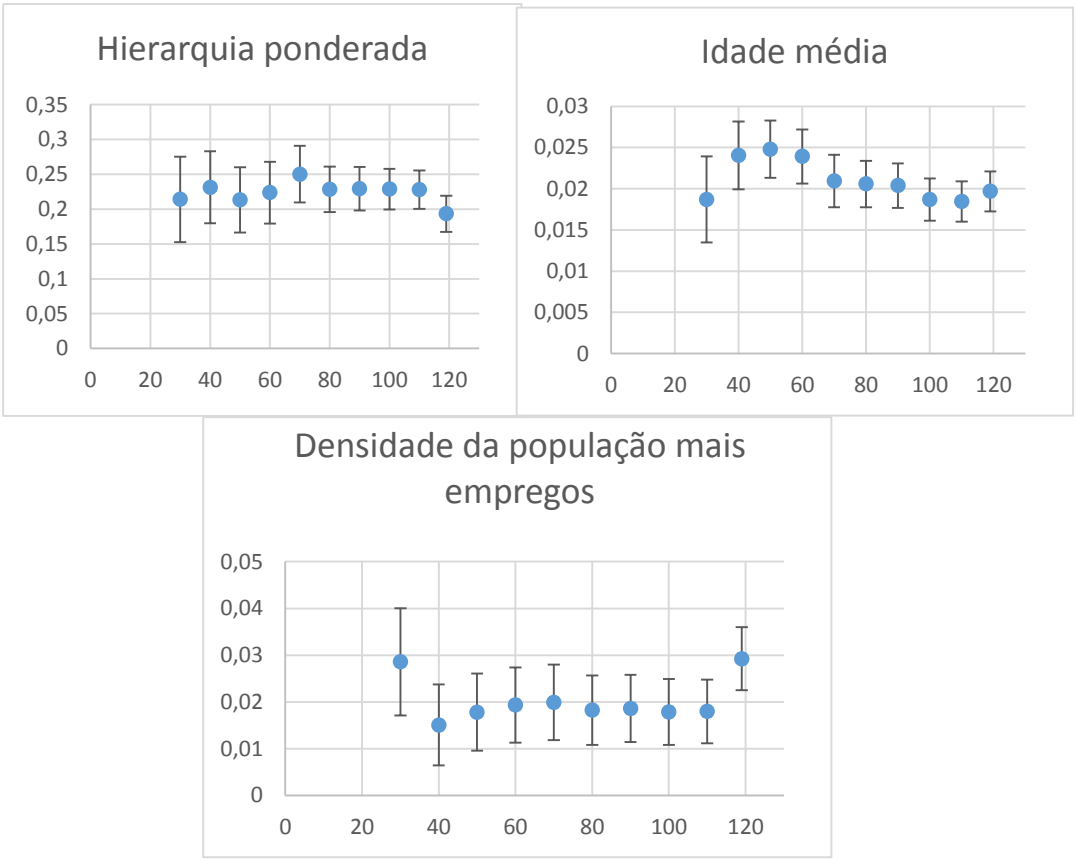


Figura 65 Gráficos da média e do erro padrão dos parâmetros da regressão da validação

Tabela 17 Comparação entre os resultados do R2, índice de Moran e teste White obtidos na calibração e na validação nos diversos níveis de agregação

| Níveis de agregação | Calibração | | | Validação | | |
|---|---|---|---|---|---|---|
| | $R^2$ | I Moran | White | $R^2$ | I Moran | White |
| **30** | 0,73 | 0,03 | 0,86 | 0,71 | 0,01 | 0,7571 |
| **40** | 0,74 | 0,13 | 0,98 | 0,71 | 0,14 | 0,7616 |
| **50** | 0,75 | 0,14 | 0,99 | 0,72 | 0,13 | 0,6401 |
| **60** | 0,74 | 0,13 | 0,99 | 0,71 | 0,1 | 0,6263 |
| **70** | 0,71 | 0,16 | 0,99 | 0,68 | 0,12 | 0,4806 |
| **80** | 0,73 | 0,16 | 0,98 | 0,7 | 0,13 | 0,5246 |
| **90** | 0,73 | 0,18 | 0,93 | 0,7 | 0,15 | 0,3468 |
| **100** | 0,72 | 0,2 | 0,86 | 0,69 | 0,15 | 0,2386 |
| **110** | 0,72 | 0,22 | 0,78 | 0,68 | 0,16 | 0,1149 |
| **119** | 0,71 | 0,2 | 0,57 | 0,68 | 0,22 | 0,0059 |

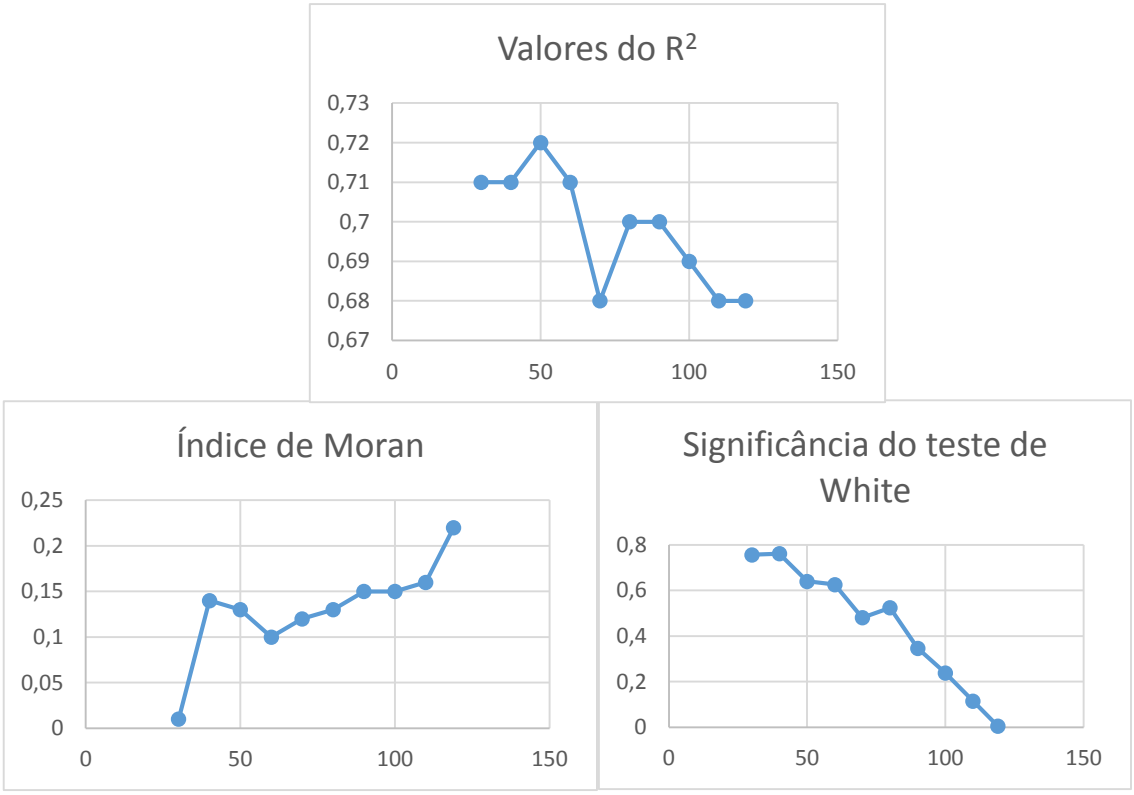


Figura 66 Gráficos de $R^2$, índice de Moran e teste de White dos resíduos da validação

Ao se verificar o valor do $R^2$, percebeu-se que houve uma queda em torno de 0,03 em todos os níveis de agregação. Os valores dos índices de Moran apresentarem resultados próximos, sendo que nos dados da validação apresentaram valores ligeiramente menores que na calibração. No entanto, foram os valores da significância do

teste de White que apresentaram as maiores diferenças, sendo que na validação obtiveram-se valores de significância consideravelmente menores que na calibração, diminuindo de forma brusca à medida em que se aumentava o nível de agregação.

De forma a comparar o comportamento das variáveis explicativas individualmente, processou-se o modelo de regressão múltipla com cada uma das variáveis individualmente. Os resultados do $R^2$, índice de Moran e teste White obtidos na calibração e na validação no nível de agregação de 119 polígonos constam na Tabela 18.

Ao se verificar quantos valores obtidos nos acidentes em 2011 estão dentro do intervalo de previsão, verificou-se que 2 valores estão abaixo do limite inferior do intervalo de previsão (Figura 67) e 14 valores estavam acima do valor máximo do intervalo de previsão (Figura 68), a um nível de confiança de 95%. Já no nível de confiança de 99%, verificou-se que nenhum dos valores encontrava-se abaixo do limite inferior do intervalo de previsão e somente um valor estava acima do valor máximo do intervalo de previsão (Figura 69). Comparando-se com o mapa da Figura 64, no qual constam as variações na densidade de acidentes entre os dados da calibração e da validação, verifica-se que os valores fora dos intervalos são aqueles situados nos intervalos com menores variações, ao contrário do que se poderia esperar. Ao se comparar com o mapa dos resíduos da regressão (Figura 46), verifica-se que 8 dos 14 valores acima do limite superior para o caso da significância de 95% encontram-se com valores de desvio-padrão acima de 1 desvio-padrão acima da média dos resíduos. Para o caso do valor situado acima do limite superior no nível de significância de 99%, o mesmo se encontra acima de 2 desvios padrão. Tal fato pode mostrar que os valores podem ter ficado fora do intervalo em parte devido ao valor empregado na calibração já se encontrar mais distante da média.

Tabela 18 Comparação entre os resultados do R2, índice de Moran e teste White das obtidos na calibração e na validação quando empregadas as variáveis explicativas individualmente com nível de agregação de 119 polígonos

| **Variáveis** | **Calibração** | | | **Validação** | | |
|---|---|---|---|---|---|---|
| | **$R^2$** | **I Moran** | **White** | **$R^2$** | **I Moran** | **White** |
| **Hierarquia ponderada** | 0,49 | 0,11 | 0,50 | 0,46 | 0,08 | 0,88 |
| **Idade média** | 0,43 | 0,47 | 0,27 | 0,41 | 0,48 | 0,19 |
| **Densidade da população mais emprego** | 0,23 | 0,007 | 0,31 | 0,21 | 0,25 | 0,0039 |

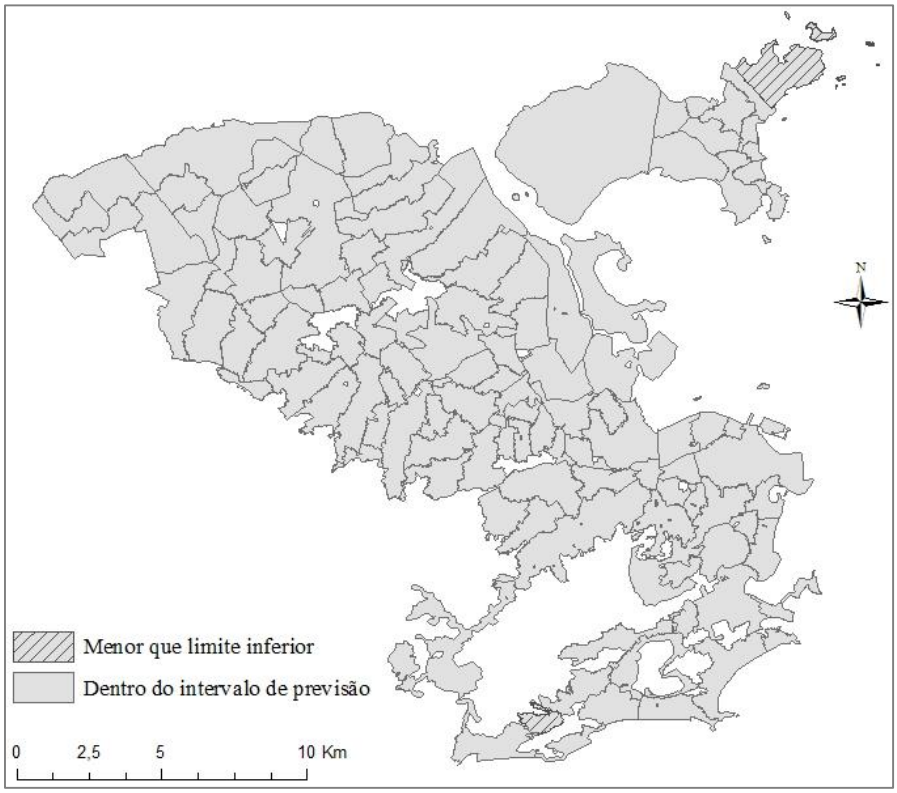


Figura 67 Bairros com valores da densidade de acidentes menor que o limite inferior do intervalo de previsão para o nível de confiança de 95%

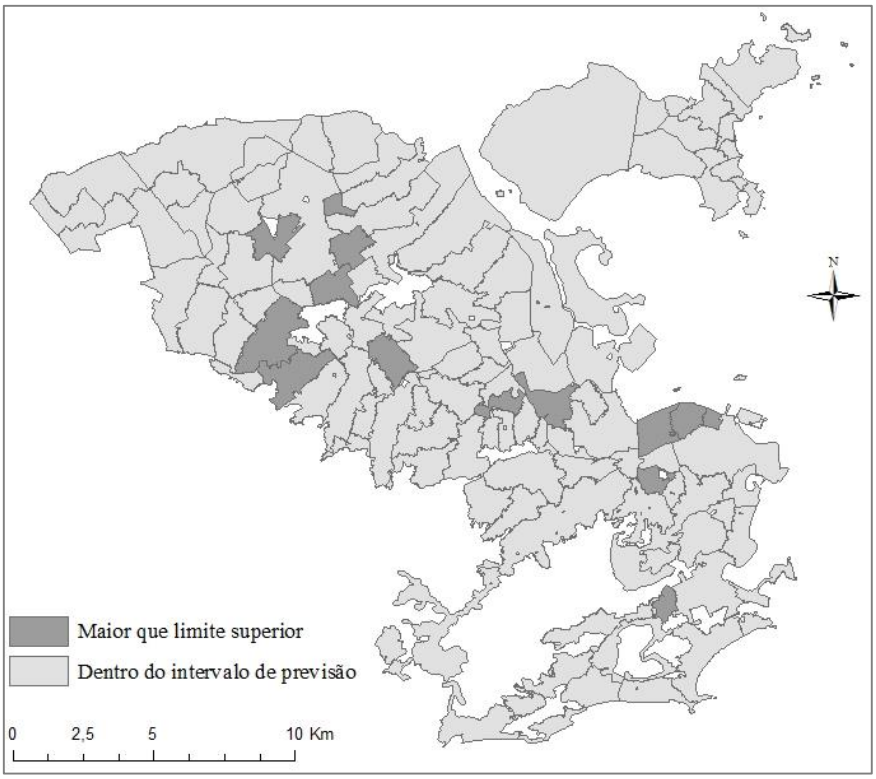


Figura 68 Bairros com valores da densidade de acidentes maiores que o limite superior do intervalo de previsão para o nível de confiança de 95%

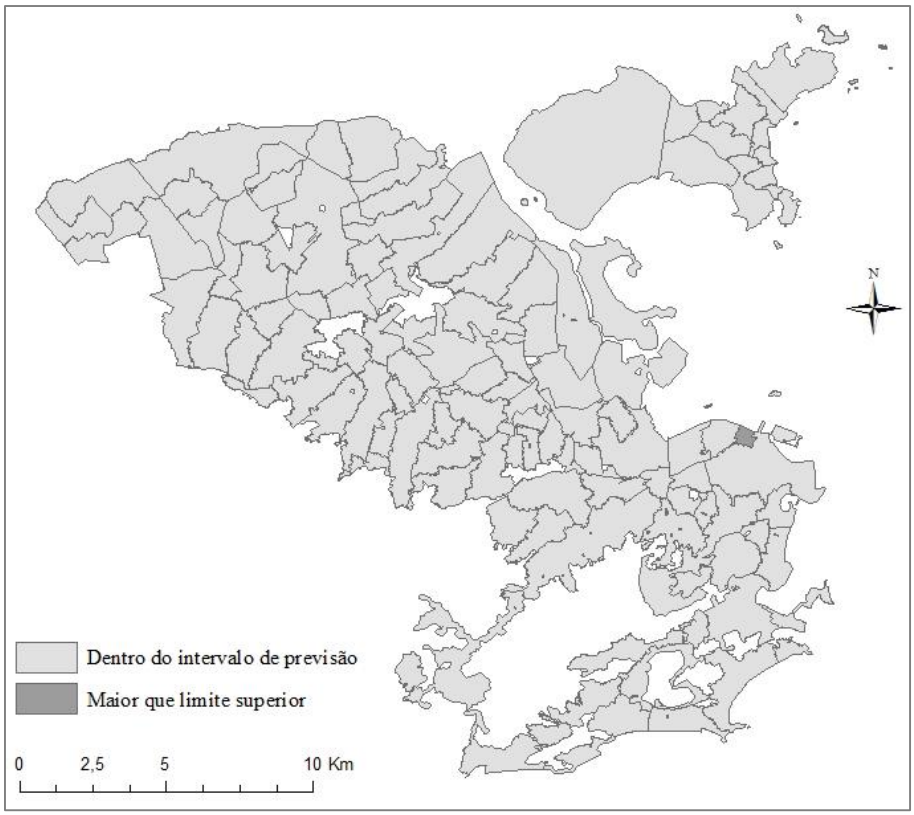


Figura 69 Bairros com valores da densidade de acidentes maiores que o limite superior do intervalo de previsão para o nível de confiança de 99%

## 5.12 Síntese dos resultados

A partir da verificação dos resultados é possível observar, no que diz respeito às hipóteses levantadas no início desta tese e para o caso da cidade do Rio de Janeiro, que:

1) A hierarquia das vias, quando ponderada pela extensão das mesmas no nível de agregação de bairro, mostrou-se como uma variável indicadora do risco de acidentes neste local. Este resultado parece coerente, tendo em vista as vias com maior hierarquia terem maior fluxo de veículos do que aquelas com menor hierarquia. Outros pesos poderiam ser testados mas aqueles adotados na pesquisa já mostram a viabilidade de se empregar tal variável;

2) Os modelos espaciais SAR e CAR não apresentam resultados consideravelmente melhores em relação aos modelos equivalentes não espaciais. Tal fato pode ter ocorrido não somente pela limitação do próprio modelo como também pela limitação da própria matriz de proximidade. O fato da região norte apresentar bairros com menor área e com menor quantidade de áreas verdes faz com que a sua configuração espacial seja diferente da região mais ao sul. Nesse sentido, o uso de um critério único de proximidade para toda

a região pode ter prejudicado o resultado dos modelos espaciais;

3) Os modelos estatísticos, quando processados em conjunto apresentaram resultados melhores do que os modelos equivalentes divididos em regimes espaciais. Tal fato pode ter ocorrido por ter havido um aumento na densidade de acidentes nos bairros na mesma medida que os indicadores socioeconômicos melhoraram. Nesse sentido, embora a cidade do Rio de Janeiro apresente indicadores demográficos e socioeconômicos heterogêneos, verificou-se que os acidentes e as variáveis explicativas mais significativas variaram na mesma direção, a qual seria em torno do sentido norte-sul da cidade. É importante mencionar que os resultados das análises visual e exploratória da variável resposta mostram indícios da existência de regimes espaciais. No entanto, ao se comparar com a distribuição de cada uma das demais variáveis explicativas, é possível ver uma certa similaridade entre as representações. Nesse sentido, a heterogeneidade espacial estaria mais presente caso as respostas das variáveis explicativas fossem diferentes em um regime em relação aos demais regimes.

Ao se analisar o teste de Chow, quando do emprego dos regimes espaciais, é possível verificar que somente a hierarquia ponderada apresenta valores não significativos, o que indica o mesmo comportamento do valor do coeficiente da variável em ambos os regimes. Já as demais variáveis, inclusive a constante, apresentam valores diferentes dos coeficientes em ambos os regimes.

Os modelos de regressão múltipla apresentaram resultados melhores do que os obtidos com os modelos lineares generalizados com distribuição binomial negativa. Tal fato pode ter ocorrido tendo em vista que as variáveis mais explicativas possuíam individualmente uma relação próxima de uma relação exponencial com a densidade de acidentes. No momento em que se fez a transformação de Box e Cox na variável dependente com vistas a atender os pressupostos da regressão múltipla, esta relação refletiu-se nos resultados obtidos da regressão.

4) A divisão da região em quantidades diferentes de áreas de agregação visando contemplar o MAUP modificou muito pouco os coeficientes dos modelos.

Conforme mencionado anteriormente, a hierarquia ponderada apresentou boa correlação com a densidade de acidentes. O fato das vias de maior hierarquia (estruturais, arteriais primárias e secundárias) apresentarem grande impacto no número de acidentes, faz com que um dado bairro apresente maior quantidade de acidentes mais pela quantidade de vias de maior hierarquia do que pela extensão das vias propriamente dito.

É possível verificar que existe boa correlação entre a densidade de acidentes e as

variáveis associadas às características socioeconômicas da população, com destaque para a densidade de estabelecimentos e densidade de empregos, bem como o percentual de moradores de favelas. A forte correlação com os dados de empregos com carteira assinada e quantidade de estabelecimentos revela o fato de que nos bairros mais pobres, principalmente nas favelas, ocorre uma baixa oferta de empregos formais e de estabelecimentos, ao contrário do que ocorre nos bairros em torno do Centro da cidade, onde reside a população com melhores indicadores socioeconômicos.

Pode-se também comentar o fato de que os bairros mais centrais e os mais ricos apresentam um fluxo de pessoas e de veículos mais intenso por mais horas do dia que os bairros mais pobres, os quais pela baixa oferta de empregos acabam por funcionar como bairros-dormitório. Além disso, no caso da cidade do Rio de Janeiro, os bairros mais valorizados estão situados próximos a grandes áreas de lazer, tais como praias, lagoa e parques, e com grande oferta de bares, restaurantes e lojas, o que faz com que aumente o fluxo de pessoas e de veículos nesses locais. A renda média per capita, apesar de ter uma certa correlação com os acidentes, apresentou resultado pior do que o da idade média talvez pelo fato de que nos bairros mais ricos alguns tipos de renda não sejam declarados ao mesmo tempo em que nos mais pobres a renda esteja incluindo o rendimento com bolsas e auxílios.

Existe correlação entre a densidade de acidentes e a variável demográfica idade média da população, reflexo da maior expectativa de vida da população nos locais de maior renda. A variável população, por sua vez, apresentou maior correlação quando testada com a variável acidentes de forma absoluta, o que mostra que nos bairros mais populosos ocorrem maior quantidade de acidentes. No entanto quando testada com a variável densidade de acidentes, a correlação caiu consideravelmente o que indica que os bairros mais populosos não são os que apresentam necessariamente maior densidade de acidentes.

Existe correlação entre a densidade de acidentes e as variáveis associadas à acessibilidade aos transportes públicos, principalmente o número de linhas de ônibus. No entanto tal variável foi retirada da modelagem pela sua alta correlação com a hierarquia ponderada. Sua alta correlação com os empregos mostra que há uma tendência de haver maior quantidade de linhas de ônibus nos bairros onde se tenha maior oferta de empregos.

Na validação do modelo, processou-se o modelo com as mesmas variáveis explicativas, alterando somente a variável resposta. Sabendo-se que houve um aumento de pouco mais de 15% no número total de acidentes empregado na modelagem e na

validação, seria de se esperar que houvesse uma certa alteração no valor do ajuste dos dados de acidentes na validação, o que foi comprovado com uma alteração no valor do $R^2$ para baixo, em torno de 0,3 em quase todos os níveis de agregação. Os valores do índice de Moran também tiveram um comportamento similar ao do modelo em todos níveis de agregação com uma tendência de diminuição no seu valor, mesmo que pequena.

No entanto, o que mais chamou a atenção foi justamente que os valores da heterogeneidade espacial nos modelos da validação, embora diminuíssem à medida em que se aumentava o número de regiões, a sua queda foi muito mais brusca que a que ocorreu no modelo da calibração. Tal fato pode ter ocorrido devido às variações mais bruscas dos acidentes ocorridos em alguns bairros, o que ficou cada vez mais explícito ao se diminuir a agregação. Quando se tinham agregações maiores, o aumento nos acidentes em um local poderia ficar encoberto, na medida em que poderia ser compensado por outros bairros em que se tinha uma diminuição no número de acidentes.

# 6. CONCLUSÕES E RECOMENDAÇÕES

A metodologia proposta mostrou-se viável com os dados de acidentes disponíveis do município do Rio de Janeiro. Traz uma contribuição quanto às variáveis utilizadas, na medida em que propõe a variável resposta densidade de acidentes, bem como as variáveis explicativas hierarquia ponderada, idade média e densidade da população mais empregos, não comumente empregadas nos estudos de análise de acidentes.

Quanto à variável hierarquia ponderada, pode ser testada com outros pesos, empregando como indicador o fluxo de veículos em algumas vias de cada hierarquia, bem como a largura ou número de faixas das mesmas.

As análises visual e exploratória, embora constem como etapas anteriores à modelagem, estão presentes em diversas etapas da metodologia, auxiliando na compreensão das variáveis e na interpretação dos resultados. O emprego de outras técnicas que as apresentadas nesta metodologia podem elucidar ainda mais a compreensão da distribuição dos acidentes e detecção de novas variáveis explicativas.

A validação dos dados aqui apresentada pode acrescentar à forma usual as análises visual e exploratória dos dados empregados, bem como a verificação do MAUP, da dependência e da heterogeneidade espaciais.

Nesta metodologia também estão contempladas as diversas especificidades dos dados geográficos, o que enriquece o estudo da distribuição espacial dos acidentes e abre a possibilidade do emprego de diversos modelos espaciais.

Por outro lado, é importante atentar para a necessidade de investigar outras variáveis explicativas de modo a que se possa obter novas variáveis correlacionadas com a variável dependente. Seria importante testar outras técnicas, tal como análise fatorial, que evitassem que variáveis explicativas com boa correlação com a variável resposta fossem eliminadas da modelagem.

É importante alertar para as limitações ocorridas nos dados de acidentes utilizados na pesquisa. A primeira seria a de não incluir os acidentes ocorridos nas vias especiais, a segunda é a dos mesmos não apresentarem informações sobre a severidade dos acidentes o que fez com que se aplicasse os modelos de frequência de acidentes. Além disso, não se pode descartar a possibilidade de haver subamostragem nos valores dos acidentes, principalmente nos locais mais pobres, tendo em vista a situação dos envolvidos nos acidentes terem maior chance de apresentarem irregularidades tais como automóvel com documentação atrasada e carteira vencida, o que pode fazer com que se evite chamar a

Polícia Militar para registrar o acidente. Além disso, nestes locais a maior parte dos automóveis não possui seguro, o que diminui o interesse do registro policial nos acidentes com somente danos materiais, pois não se tem a necessidade de entregar este documento na seguradora do veículo.

Embora o valor da dependência seja considerado relativamente baixo (em torno de 0,2), faz-se necessário buscar modelos que diminuam tal dependência. Tal efeito pode ter ocorrido tanto pela má especificação do modelo, como pela omissão de alguma variável explicativa que capte a dependência espacial.

Embora o modelo tenha obtido um bom resultado para o período testado, sua eficácia para dados futuros do mesmo local parece limitada, tendo em vista que o mesmo está baseado em variáveis que não aumentam no mesmo percentual que o dos acidentes. Nesse sentido, a hierarquia ponderada embora seja uma variável indicadora do risco de acidentes neste dado momento com o passar do tempo poderá não continuar sendo. Por outro lado, para que se tenha condições de utilizar as mesmas variáveis explicativas para um período diferente no mesmo local, seria interessante que fosse acrescentado ao modelo uma variável que captasse essa tendência de aumento nos acidentes. A inserção de outras variáveis mais sensíveis ao aumento do número de acidentes como, por exemplo, a frota de veículos, pode ser uma forma de melhorar o ajuste do modelo no mesmo local em um outro momento.

O objetivo desta tese, tanto quanto encontrar o melhor modelo dentre os disponíveis, foi a de apresentar uma metodologia para trabalhar com as variáveis quando as mesmas são dados geográficos, considerando suas especificidades e oferecendo a possibilidade de analisar os dados e os resultados de forma espacial, deixando espaço para que se empregue diferentes técnicas de visualização e análise exploratória e se teste outros modelos não espaciais e espaciais. Desse modo, tão importante quanto identificar e quantificar algumas variáveis significativas, tem-se o caminho para tratar as mesmas.

## REFERÊNCIAS BIBLIOGRÁFICAS


ABRACICLO. **Anuário da Indústria Brasileira de Duas Rodas**, São Paulo, 2012. Disponível em <http://www.abraciclo.com.br>. Acesso em 25 out 2012.

ABNT, NBR 12898: **Relatório de Acidente de Trânsito (RAT).** Rio de Janeiro, 1993.

ALMEIDA, R.L.F. **Epidemiologia dos acidentes de trânsito em Fortaleza no período de 2004 a 2008**, 2011. 158f. Tese de doutorado. (Doutorado em Saúde Coletiva), Coordenação do Curso de Pós-Graduação em Saúde Coletiva, Universidade Estadual do Ceará, Universidade Federal do Ceará, Fortaleza, 2011.

ALMEIDA, E. **Econometria Espacial Aplicada**. Editora Alínea. Campinas, SP. 2012.

ALVES, P. **Correlação entre acidentes de trânsito, uso e ocupação do solo, polos geradores de viagens e população na cidade de Uberlândia**. 184f. Dissertação. (Mestrado em Engenharia Urbana), Pós-Graduação em Engenharia Urbana, Centro de Ciências Exatas e de Tecnologia, Universidade Federal de São Carlos, São Carlos, SP, 2011.

AGUERO-VALVERDE, J. Full Bayes Poisson gamma, Poisson lognormal, and zero inflated random effects models: comparing the precision of crash frequency estimates. **Accident Analysis and Prevention**, v. 50, pp.289-297, 2013.

AGUERO-VALVERDE, J., JOVANIS, P.P. Spatial correlation in multilevel crash frequency models: effects of different neighboring structures. **Transportation Research Record** v. 2165, pp. 21–33, 2010.

AGUERO-VALVERDE, J., JOVANIS, P.P. Spatial analysis of fatal and injury crashes in Pennsylvania. **Analysis and Prevention**, v. 38, 618–625, 2006.

ANASTASOPOULOS, P.C, MANNERING, F.L. An empirical assessment of fixed and random parameter logit models using crash and non crash specific injury data. **Accident Analysis and Prevention**, v. 43, pp.1140-1147, 2011.

ANDERSON, T. Kernel density estimation and K-means clustering to profile road

accident hotspots. **Accident Analysis and Prevention**, v. 41, pp. 359 – 364, 2009.

ANFAVEA. **Anuário Estatístico da Indústria Automobilística Brasileira**, São Paulo, 2012. Disponível em <http://www.anfavea.com.br/anuario.html>. Acesso em 25 out 2012.

ANSELIN, L. (2005). **Exploring spatial data with geoda: a workbook.** Santa Barbara, US. Disponível em <http://www.csiss.org>. Acesso em: 31 out 2014.

ANSELIN, L., FLORAX, R.J.G.M., REY, S.J. (ed.). **Advances in spatial econometrics**, Nova York: Springer, 2004.

ANSELIN, L. (2002). **Mapping and analysis for spatial social science.** Santa Barbara, US. Disponível em http://www.csiss.org.Acesso em: 12 jul 2012.

ANSELIN, L. **The Moran scatterplot as an ESDA tool to assess local instability in spatial association.** In: Spatial Analytical Perspectives on GIS in Environmental and Socio-Economic Sciences. London: Taylor and Francis, pp. 111-125, 1996.

ANSELIN, L. Some robust approaches to testing and estimation in spatial econometrics. **Regional Science and Urban Economics**, v. 20, pp. 141-163, 1990.

ANSELIN, L. **Spatial Econometrics: methods and models**. Boston: Kluwer Academic, 1988.

ARAUJO, G. P.., BRAGA, M.G.C. Methodology for the quantitative evaluation of pedestrian crossings at road junctions with traffic lights. **Transportation**, v. 35, pp. 539-557, 2008.

AZIZ, H.M.A., UKKUSURI, S.V., HASAN, S. Exploring the determinants of pedestrian–vehicle crash severity in New York City. **Accident Analysis and Prevention**, v. 50, pp. 1298-1309, 2013.

BAILEY, T.C., GATRELL, A.C**. Interactive Spatial Data Analysis**. John Wiley & Sons, New York, 413 pp. 1995.

BARBOSA, H., CUNTO, F., BEZERRA, B., NODARI, C. Safety performance models for urban intersections in Brazil. **Accident Analysis and Prevention**, v. 70, pp. 258-266, 2014.

BASTOS, J.T. **Geografia da mortalidade no trânsito no Brasil**. 2011. 150f. Dissertação. (Mestrado em Engenharia de Transportes), Pós-Graduação em Engenharia de Transportes, Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, SP, 2011.

BERTIN, J. **Sémiologie Graphique: les diagrammes, les réseaux, les cartes**. Monton & Gauthier-Villars, Paris, 1967.

BLAZQUEZ, C. A., CELIS, M.S. A spatial and temporal analysis of child pedestrian crashes in Santiago, Chile, **Accident Analysis and Prevention**, v. 50, pp. 304-311, 2013.

BOFFO, G. H. **Formatos e técnicas de modelos de previsão de acidentes de trânsito**. Dissertação. (Mestrado em Engenharia de Produção), Pós-Graduação em Engenharia de Produção, Escola de Engenharia da UFRGS, Universidade Federal do Rio Grande do Sul, 2011.

BOX, G.E.P., COX, D. R. An analysis of transformations. **J.R. Stat. Soc**. *B* , v.26, pp.211-252, 1964.

BRAGA, M. G. C.; RIBEIRO, S. C.; FERREIRA, M. M. **Envolvimento em acidentes e exposição ao tráfego: estudo de caso para a cidade do Rio de Janeiro.** In: III Rio de Transportes, Rio de Janeiro, 2005.

BURROUGH, P. A. **Principles of Geographical Information Systems for Land Resources Assessment**. Oxford: Clarendon Press. 1986.

CARDOSO, G. **Modelos para previsão de acidentes de trânsito em vias arteriais urbanas**, 2006. 289f. Tese de doutorado. (Doutorado em Engenharia de Produção), Pós-Graduação em Engenharia de Produção, Escola de Engenharia da UFRGS, Universidade Federal do Rio Grande do Sul, 2006.

CASETTI, E. Generating models by the expansion method: applications to the investigation of fertility development relations. **Modeling and Simulation**, v.13, pp. 961-966, 1972.

CASTIGLIONE, L. H. G. **Uma viagem epistemológica ao geoprocessamento**. 283p. Dissertação. (Mestrado em Estudos Populacionais e Pesquisas Sociais), Escola Nacional de Ciências Estatísticas, 2003.

CERVERO, R., RADISCH, C. Travel choices in pedestrian versus automobile oriented neighborhoods. **Transport Policy**, v. 3, pp. 127– 141, 1996.

CLIFTON, K.J., KREAMER-FULTS, K. An examination of the environmental attributes associated with pedestrian-vehicular crashes near public schools. **Accident Analysis and Prevention**, v. 39, pp. 708-715, 2007.

CORDEIRO, M.G., DEMÉTRIO, C.G.B. **Modelos Lineares Generalizados**. Minicurso para o 12º SEAGRO e a 52ª Reunião Anual da RBRAS, Universidade Federal de Santa Maria, Santa Maria, 2008.

COTTRILL,C.D., THAKURIAH, P.V. Evaluating pedestrian crashes in areas with high low-income or minority populations. **Accident Analysis and Prevention**, v. 42, pp. 1718-1728, 2010.

COWEN, D.J. Computer mapping in GIS: implications for applied geography. Papers and Proceedings of the Applied Geography Conferences, 10, 43. 1987

CUNTO, F.J.C., CASTRO NETO, M.M., BARREIRA, D.S. **Modelos de previsão de acidentes de trânsito em interseções semaforizadas de Fortaleza**. In: *XXV Congresso de Pesquisa e Ensino em Transportes*. Belo Horizonte-MG, ANPET(Ed.), 2011.

DENATRAN. **Manual de identificação, análise e tratamento de pontos negros, 2ª edição, coleção serviços de engenharia**. Departamento Nacional de Trânsito, Brasília, DF, 1987.

DENATRAN. **Frota de veículos em 2001 e 2011**, Brasília, 2012. Disponível em <http://www.denatran.gov.br>. Acesso em 27 out 2012.

DENT, B.D. **Principles of Thematic Map Design**. Addison-Wesley Publishing Company, 398 pp. 1985.

DISSANAYAKE, D., ARYAIJA, J., WEDAGAMA, D.M.P. Modelling the effects of land use and temporal factors on child pedestrian casualties. **Accident Analysis**

**and Prevention**, v. 41, pp. 1016-1024, 2009.

DONALDSON, A.E., COOK, L.J., HUTCHINGS, C.B., DEAN, J.M. Crossing county lines: the impact of crash location and drivers residence on motor vehicle crash fatality. **Accident Analysis and Prevention**, v. 38, pp. 723-727, 2006.

DRUCK, S., CARVALHO, M.S., CÂMARA, G., MONTEIRO, A.V.M. (eds). **Análise Espacial de Dados Geográficos**. EMBRAPA, Brasília, DF, 2004.

EDWARDS, P., GREEN, J., ROBERTS, I., LUTCHMUN, S. Deaths from injury in children and employment status in family: analysis of trends in class specific death rates. **British Medical Journal 333 (July)**, pp. 119–121, 2006.

ELIAS, W., SHIFTAN, Y. Analyzing and modeling risk exposure of pedestrian children to involvement in car crashes. **Accident Analysis and Prevention**, v. 62, pp. 397-405., 2014.

FERREIRA, F.F. Fatores de risco envolvendo motocicletas em vias urbanas: a percepção dos condutores profissionais. 91f. Dissertação. (Mestrado em Engenharia de Produção), Pós-Graduação em Engenharia de Produção, Escola de Engenharia da UFRGS, Universidade Federal do Rio Grande do Sul, 2009.

FHWA. **Guidebook on methods to estimate non-motorized travel: supporting documentation, Federal Highway of Transportation**. U.S. Department of Transportation. 1999.

FLAHAUT, B., MOUCHART, M., SAN MARTIN, B., THOMAS, I. The local spatial autocorrelation and the kernel method for identifying black zones: a comparative approach. **Accident Analysis and Prevention**, v. 35, pp. 991 – 1004, 2003.

FOTHERINGHAM, A.S., BRUNSDON, C., CHARLTON, M.E. **Quantitative Geography: Perspectives on Spatial Data Analysis**. Sage Publications, London, 2000.

GILES-CORTI, B., WOOD, G., PIKORA, T., LEARNIHAN, V., BULSARA, M., VAN NIEL, K., TIMPERIO, A., MCCORMACK, G.; VILLANUEVA, K. School site and the potential to walk to school: the impact of street connectivity and traffic exposure in school neighborhoods. **Health&Place**, v. 17, pp. 545-550, 2011.

GOLD, P. A. **Segurança de trânsito: aplicação de engenharia para reduzir acidentes**. Banco Interamericano de Desenvolvimento, Washington, 1998.

GOMES, L.P., MELO, E.C.P. Distribuição da mortalidade por acidentes de trânsito no município do Rio de Janeiro. **Escola Anna Nery Revista de Enfermagem**, v. 11, n. 2, pp. 289-295, 2006.

GOODCHILD, M.F. **The validity and usefulness of laws in geographic information science and geography**. In: Annals of the Association of American Geographers, v. 94, n. 2, pp. 300-303, 2004.

GOODCHILD, M.F. A spatial analytical perspective on geographical information systems. **International Journal of Geographical Information Systems**, v. 1, n. 4, pp. 327-334, 1987.

GRAHAM, D.J., GLAISTER, S., ANDERSON, R. The effects of area deprivation on the incidence of child and adult pedestrian casualties in England. **Accident Analysis and Prevention**, v. 37, pp. 125-135, 2005.

GRAHAM, D.J., GLAISTER, S. Spatial Variation in road pedestrian casualties: the role of urban scale, density and land-use mix. **Urban Studies**, v. 40, n. 8, pp. 1591-1607, 2003.

GREEN, J., MUIR, H., MAHER, M. Child pedestrian casualties and deprivation. **Accident Analysis and Prevention**, v. 43, pp. 714-723, 2011.

GUJARATI, D.N. **Basic Econometrics**, 4ª Edição,. Editora McGraw Hill, Boston, 2003.

GUO, D., WANG, H. Automatic region building for spatial analysis. **Transations in GIS**, v. 15, pp. 29-45 (s1), 2011.

GUO, D. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). **International Journal of Geographical Information Science**, v. 22, n. 7, pp. 801-823, 2008.

HA, H.H., THILL, J.C. Analysis of traffic hazard intensity: a spatial epidemiology case study of urban pedestrians. **Computers, Environment and Urban Systems**, v. 35, pp. 230-240, 2011.

HAINING, R. **Spatial Data Analysis: theory and practice**. Cambridge University Press, Cambridge, 2003.

HOLZ, R.F., KORZENOWSKI, A., NODARI, C.T., TEN CARTEN, C.S., LINDAU, L.A., **Modelagem de acidentes envolvendo motociclistas em Porto Alegre**, In: *XXV Congresso de Pesquisa e Ensino em Transportes*. Belo Horizonte-MG, ANPET(Ed.), 2011.

HUANG, H., ABDEL-ATY, M. Multilevel data and Bayesian analysis in traffic safety. **Accident Analysis and Prevention**, v. 42 (6), pp. 1556-1565, 2010.

IPEA, DENATRAN, ANTP. **Impactos sociais e econômicos dos acidentes de trânsito nas rodovias brasileiras: relatório executivo**, Brasília, 2006.

IPEA. **Impactos sociais e econômicos dos acidentes de trânsito nas aglomerações urbana: relatório executivo**. Brasília, 2003.

JOHNSON, G.D., LU, X. Neighborhood-level built environment and social characteristics associated with serious childhood motor vehicle occupant injuries. **Health & Place**, v. 17, pp. 902-910, 2011.

KELEJIAN, H.H., ROBINSON, D.P. Spatial autocorrelation: a new computationally simple test with an application to per capita county policy expenditures. **Regional Science and Urban Economics**, v.22, pp. 317-331, 1992.

KELLY, C.E., TIGHT, M.R., HODGSON, F.C., PAGE, M.W. A comparison of tree methods for assessing the walkability of the pedestrian environment. **Journal of Transport Geography**, v. 19, pp. 1500.1508, 2011.

KIM, K., YAMASHITA, E.Y. **The influence of land use, population, employment and economic activity on accidents**. In: 85th Transportation Research Record Annual Meeting, 20p, 2006.

KINGHAM, S., SABEL, C.E., BARTIE, P. The impact of the 'school run' on road traffic accidents: a spatio-temporal analysis, **Journal of Transport Geography**, v. 19, pp. 705-711, 2011.

KUHLMANN, A.K.S., BRETT, J., THOMAS, D., SAIN, S.R. Environmental characteristics associated with pedestrian-motor vehicle collision in Denver.

**American Journal of Public Health**, v. 99, n. 9, pp. 1632-1637, 2009.

LANDIS, B.W., VATTIKUTI, V.R., OTTENBERG, R.M., MCLEOD, D.S.; GUTTENPLAN, M. Modeling the roadside walking environment: a pedestrian level of service. **Transportation Research Record**, v. 1773, p 82-88, 2001.

LASCALA, E.A., GUENEWALD, P.J., JOHNSON, F.W. An ecological study of the locations of schools and child pedestrian injury collisions. **Accident Analysis and Prevention**, v. 36, pp. 569-576, 2004.

LASSARE, S., PAPADIMITRIOU, E., YANNIS, G.,GOLIAS, J. Measuring accident risk exposure for pedestrians in different micro-environments. **Accident Analysis and Prevention**, v. 39, n. 6, pp. 1226–1238, 2012a.

LASSARE, S., BONNET, E., BODIN, F., PAPADIMITRIOU, E., YANNIS, G., GOLIAS, J. A GIS-based methodology for identifying pedestrians' crossing patterns. **Computers, Environment an Urban Systems**, v. 36, pp. 321-330, 2012b.

LI, Z., WANG, W., LIU, P., BIGHAM, J.M., RAGLAND, D.R. Using geographically weighted Poisson regression for county-level crash modeling in California. **Safety Science**, v. 58, pp. 89-97, 2013.

LICAJ, I., HADDAK, M., POCHET, P., CHIRON, M. Contextual deprivation, daily travel and road traffic injuries among the young in the Rhône Département (France). **Accident Analysis and Prevention**, v. 43, pp. 1617-1623, 2011.

LORD, D., MANNERING, F. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. **Transportation Research Part A**, v. 44, pp.291-305, 2010.

MADALOZO, H.C., DYMINSKI, A.S. **Análise de curvas horizontais de rodovias para melhoramento de projeto e operação utilizando redes neurais artificiais**. *In: XXIII Congresso de Pesquisa e Ensino em Transportes.* Florianópolis-SC, ANPET(Ed.), 2009.

MAGUIRE, D, J., GOODCHILD, M., RHING, D.W . **Geographic Information Systems: Principles and Applications.** Ed. Longman/Wiley, 1991.

MAIA, P.B., AIDAR, T. **Análise dos resultados do pareamento dos Boletins de Ocorrência e das declarações de óbitos referentes aos acidentes de trânsito do município de São Paulo – 2003-2004**. In: XVII Encontro Nacional de Estudos Populacionais. Caxambu-MG, ABEP(Ed.), 2010.

MÂNICA, A.G. **Modelo de previsão de acidentes rodoviários envolvendo motocicletas**, 2007. 177f. Tese de doutorado. (Doutorado em Engenharia de Produção), Pós-Graduação em Engenharia de Produção, Escola de Engenharia da UFRGS, Universidade Federal do Rio Grande do Sul, 2007.

MARSHALL, S. **Streets & Patterns**. Spon Press, New York, 2005.

MARSHALL, W.E., GARRICK, N.W. The Effect of Street Network Design on Walking and Biking. **Journal of the Transportation Research Board**, v. 2198, pp.103-115, 2010a.

MARSHALL, W. E., GARRICK, N.W. Does street network design affect traffic safety? **Accident Analysis and Prevention**, v. 43, pp. 769-781, 2011.

MAC EACHREN, A.M. **Visualization in Modern Cartography**. Pergamon Press, 1994.

MALTA, D.C., MASCARENHAS, M.D.M., BERNAL, R.T.I., SILVA, M.M.A., PEREIRA, C.A., MINAYO, M.C.S., MORAIS NETO, O.L. Análise das ocorrências das lesões no trânsito e fatores relacionados segundo resultados da Pesquisa Nacional por Amostra de Domicílios (PNAD) – Brasil, 2008**. Ciência Saúde Coletiva**, v.16, n.9, pp. 3679-3687, 2011.

MIRANDA-MORENO, L.F., MORENCY. P., EL-GENEIDY, A.M. The link between built environment, pedestrian activity and pedestrian - vehicle collision occurrence at signalized intersections. **Accident Analysis and Prevention**, v. 43 (5, pp. 1624-1634, 2011.

MONICO, J. F. G. **Posicionamento pelo NAVSTAR-GPS: descrição, fundamentos e aplicações**. Editora Unesp, 291 pp. 2000.

MONTGOMERY, D.C., RUNGER, G,C**. Estatística Aplicada e Probabilidade para Engenheiros**, 5ª Edição, Editora LTC, 2012.

MORAIS NETO , O.L., MONTENEGRO, M.M.S., MONTEIRO, R.P., SIQUEIRA JÚNIOR, J.B., SILVA, M.M.A., LIMA, C.M., MIRANDA, L.O.M., MALTA, D.C., SILVA JUNIOR, J.B. Mortalidade por acidentes de transporte terrestre no Brasil na última década: tendência e aglomerados de risco. **Ciência e Saúde Coletiva**, v. 17, n. 9, pp.2223-2236, 2012.

NODARI, C.T. **Método de Avaliação da Segurança Potencial de Segmentos Rodoviários Rurais de Pits Simples**, 2003. 221f. Tese. (Doutorado em Engenharia de Produção), Pós-Graduação em Engenharia de Produção, Escola de Engenharia da UFRGS, Universidade Federal do Rio Grande do Sul, 2003.

NOLAND, R.B., KLEIN, N.J., TULACH, N.K. Do lower income areas have more pedestrian casualities? **Accident Analysis and Prevention**, v.59, pp.337-345, 2013.

NOLAND, R. B., QUDDUS, M.A. A spatially disaggregate analysis of road casualties in England. **Accident Analysis and Prevention**, v. 36, pp. 973-984, 2004.

OPENSHAW, S. The Modifiable Areal Unit Problem. **Concepts and Techniques in Modern Geography**, v. 38, 1984.

PANTER, J.R., JONES, A.P., VAN SLUIJS, E.M.F. Neighborhood, route, and school environments and children´s active commuting. **American Journal of Preventive Medicine**, v. 38, n. 3, pp. 268-278, 2010.

PAPADIMITRIOU, E. Theory and models of pedestrian crossing behaviour along urban trips. **Transportation Research Part F,** v. 15, pp. 75-94, 2012.

PETERSON, M. P. **Spatial Visualization through Cartographic Animation: Theory and Practice.** Omaha, University of Nebraska, 1994.

PETRITSCH, T.A., HUANG, H.F., LANDIS, B.W. **Midblock pedestrian and pathway crossings of roadways.** In:86th Transportation Research Record Annual Meeting, 25p, 2007.

PULUGURTHA, S.S., DUDDU, V.R., KOTAGIRL, Y. Traffic analysis zone level crash estimation models based on land use characteristics. **Accident Analysis and Prevention** , v.50, pp. 678–687, 2013.

PULUGURTHA, S.S., SAMBHARA, V.R. Pedestrian crash estimation models for signalized intersections. **Accident Analysis and Prevention**, v. 43, pp. 439-446, 2011.

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

QUEIROZ, M.P. **Análise espacial dos acidentes de trânsito do município de Fortaleza**. 141f. Dissertação. (Mestrado em Engenharia de Transportes), Mestrado em Engenharia de Transportes, Universidade Federal do Ceará, Fortaleza, CE, 2003.

QUEIROZ, M.P., LOUREIRO, C.F.G, YAMASHITA, Y. Metodologia da análise espacial para identificação de locais críticos considerando a severidade dos acidentes de transito", **Transportes**, v. 12, pp. 15-28, 2004a.

QUEIROZ, M.P., LOUREIRO, C.F.G., CUNTO, F.J.C. **Georeferenciamento do Sistema de informações de Acidentes de Trânsito de Fortaleza (SIAT-FOR): aperfeiçoamento e vantagens.** In: XVIII Congresso de Pesquisa e Ensino em Transportes. Florianópolis-SC, ANPET(Ed.), 2004b.

RAFORD, N., RAGLAND, D.R. **Space Syntax: An Innovative Pedestrian Volume Modeling Tool for Pedestrian Safety**. Institute of Transportation Studies. U.C.Berkeley Traffic Safety Center, 2003.

ROCHA, M. M., NASSI, C. D. **Modelagem estatística dos acidentes de trânsito na cidade do Rio de Janeiro com emprego de Sistema de Informações Geográficas**. In: Congresso Panamericano de Engenharia de Trânsito, Transporte e Logística - PANAM, Santiago, 2012a.

ROCHA, M. M, NASSI, C.D. **Análise estatística e da distribuição espacial dos acidentes de trânsito na zona Sul do Rio de Janeiro**. In: Congresso Luso Brasileiro para o Planejamento Urbano Regional Integrado e Sustentável – PLURIS. Brasília. Planejamento Urbano Regional Integrado e Sustentável, 2012b.

SANTOS, L., RAIA JUNIOR, A.A. Distribuição de acidentes de trânsito em São Carlos: identificação de tendências de deslocamento através da técnica de elipse de desvio padrão. **Caminhos da Geografia**, v.7. n. 18, pp.134-145, 2006.

SAVOLAINEN, P. T., MANNERING, F. L., LORD, D., QUDDUS, M.A. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. **Accident Analysis and Prevention,** v. 43, pp.1666-1676, 2011.

SCRIBTER, R.A., MACKINNON, D.P., DWYER, J.H. Alcohol outlet density and motor vehicle crashes in Los Angeles County Cities. **J. Stud. Alcohol**, v.55, 1994.

SIDDIQUI, C., ABDEL-ATY, M. Nature of modeling boundary pedestrian crashes of zones. **Journal of the Transportation Research Board**, v. 2239, pp. 31-40, 2012

SIDDIQUI, C., ABDEL-ATY, M., CHOI, K. Microscopic spatial analysis of pedestrian and bycicle crashes. **Accident Analysis and Prevention**, v. 45, pp. 382-391, 2012.

SIG Floresta (2012). Disponível em http://sigfloresta.rio.rj.gov.br/ Acesso em 21 Jan 15.

SILVA, K.C.R. **Aplicação do modelo de previsão de acidentes do HSM em rodovias de pista simples do estado de São Paulo**. 95f. Dissertação. (Mestrado em Engenharia de Transportes), Pós-graduação em Engenharia de Transportes, Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, SP, 2011.

SILVA, M.O, CARVALHO, G. L. P. N., VERSIANI, M. H., BASTOS JUNIOR, C. S., REGO, H. R. S.; MARCELLINO, I. S. 2011. **Características e evolução recente do emprego e economia carioca e metropolitana**. Coleção Estudos Cariocas nº 20110401. Rio de Janeiro, RJ. Disponível em <http://portalgeo.rio.rj.gov.br/estudoscariocas/download/2423_Caracteristicas_e_evolucao_recente_do_emprego_e_da_economia_carioca_e_metropolitana_2.pdf>. Acesso em 29 jan 2015.

SOARES, A.J. **Análise de autocorrelação em redes aplicada ao caso de acidentes urbanos de trânsito**. 138p. Dissertação. (Mestrado em Engenharia Civil), Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, SP, 2007.

SOUZA, G.A. Georreferenciamento de acidentes de trânsito: uma discussão

metodológica. **ACTA Geográfica**, Ed. Esp. Cidades na Amazônia Brasileira, pp. 31-40, 2011.

SOUZA, G.A. **Espacialidade urbana, circulação e acidentes de trânsito: o caso de Manaus – AM (2000 a 2006)**. 139f. Tese de doutorado. (Doutorado em Engenharia de Transportes), Programa de Engenharia de Transportes, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, 2009.

STPP. Mean Streets 2002, Surface Transportation Policy Project. 2002. Disponível em <http://www.transact.org/PDFs/ms2002/MeanStreets2002.pdf>. Acesso em 12 set 2012.

THAKURIAH, P.V., METAXATOS, P., LIN, J., JENSEN, E., An examination of factors affecting propensities to use bicycle and pedestrian facilities in suburban locations. **Transportation Research Part D**, v. 17, pp. 341-348, 2012.

THAKURIAH, P.V., COTRILL, D.C. **Evaluating pedestrian risk in environmental justice areas**. In: 87th Transportation Research Record Annual Meeting, 17p, 2008.

TOBLER, W. R. 1970. A computer movie simulating urban growth in the Detroit region. **Economic Geography**, v.46, pp. 234–40, 1970.

UKKUSURI, S., MIRANDA-MORENO, L.F., RAMADURAI, G., ISA-TAVAREZ, J. The role of built environment on pedestrian crash frequency. **Safety Science**, v. 50, pp. 1141–1151, 2012.

WASHINGTON, S., VAN SCHALWYK, I., MITRA, S., MEYER, M., DUMBAUGH, E., ZOLL, M. Incorporating Safety into Long-Range Transportation Planning, 2006. NCHRP Report 546. 2006 .Disponível em: <http://onlinepubs.trb.org/onlinepubs/nchrp/nchrprpt546.pdfhttp://onlinepubs.trb.org/onlinepubs/nchrp/nchrprpt 546.pdf >. Acesso em 04 Jun 2011.

WANG, C., QUDDUS, M.A., ISON, S.G. The effect of traffic and road characteristics on road safety: a review and future research direction. **Safety Science**, v.57, pp.264-275, 2013.

WANG, C., QUDDUS, M.A., RYLEY, T., ENOCH, M., DAVISON, L. **Spatial models in transport: a review and assessment of methodological issues**. In:

91th Annual Meeting of the Transportation Research Board, Washington, DC, 2011.

WANG, Y., KOCHELMAN,K.M. A Poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts crossneighborhoods. . **Accident Analysis and Prevention**, v. 60, pp.71-84, 2013.

WHITE, H. Heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. **Econometrica**, v. 48, pp. 817-838, 1980.

WIER, M. WEINTRAUB, J., HUMPHREYS, E., SETO, E., BHATIA, R. An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning. **Accident Analysis and Prevention Journal** 41 (1), pp. 137–145, 2009.

WHO. **Global Status Report on road safety: time for action**, World Health Organization, Genebra, 2009.

XU, P., HUANG, H. Modeling crash spatial heterogeneity: Random parameter versus geographically weighting. **Accident Analysis and Prevention**, v. 75, pp.16-25, 2015.

XU, P. HUANG, H., DONG, N., ABDEL-ATY, M. Sensitivity analysis in the context of regional safety modeling: identifying and assessing the modifiable areal unit problem. **Accident Analysis and Prevention**, v. 70, p.110-120, 2014.

YANNIS, G., PAPADIMITRIOU, E., ANTONIOU, C. Impact of enforcement on traffic accidents and fatalities: a multivariate multilevel analysis. **Safety Science**, v. 46, p. 738-750, 2008.

ZHANG, Y., BIGHAM, J.; LI, Z.; RAGGLAND, D. **Associations between road network connectivity and pedestrian-byciclist accidents**. In: 91th Transportation Research Record Annual Meeting, 18p, 2013.