Rafael Ornellas Barbosa Pereira Pigozzo

Análise de roubos no município do Rio de Janeiro via modelos aditivos generalizados

Niterói - RJ, Brasil

15 de dezembro de 2022

Universidade Federal Fluminense

Rafael Ornellas Barbosa Pereira Pigozzo

Análise de roubos no município do Rio de Janeiro via modelos aditivos generalizados

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador: Prof. Dr. Rafael Santos Erbisti

Co-Orientador: Prof. Dr. Jony Arrais Pinto Junior

Niterói - RJ, Brasil

15 de dezembro de 2022

Universidade Federal Fluminense

Rafael Ornellas Barbosa Pereira Pigozzo

Análise de roubos no município do Rio de Janeiro via modelos aditivos generalizados

Monografia de Projeto Final de Graduação sob o título *"Análise de roubos no município do Rio de Janeiro via modelos aditivos generalizados"*, defendida por Rafael Ornellas Barbosa Pereira Pigozzo e aprovada em 15 de dezembro de 2022, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Prof. Dr. Rafael Santos Erbisti
Departamento de Estatística – UFF

Prof. Dr. Jony Arrais Pinto Junior
Departamento de Estatística – UFF

Prof. Dr. Luis Guillermo Coca Velarde
Departamento de Estatística – UFF

Profa. Dra. Renata Souza Bueno
ENCE\IBGE

Niterói, 15 de dezembro de 2022

Ficha catalográfica automática - SDC/BIME
Gerada com informações fornecidas pelo autor

P633a Pigozzo, Rafael Ornellas Barbosa Pereira
Análise de roubos no município do Rio de Janeiro via modelos aditivos generalizados / Rafael Ornellas Barbosa Pereira Pigozzo. - 2022.
63 f.: il.

Orientador: Rafael Santos Erbisti.
Coorientador: Jony Arrais Pinto Junior.
Trabalho de Conclusão de Curso (graduação)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2022.

1. Modelos aditivos generalizados. 2. Splines. 3. Roubos. 4. Rio de Janeiro. 5. Produção intelectual. I. Erbisti, Rafael Santos, orientador. II. Pinto Junior, Jony Arrais, coorientador. III. Universidade Federal Fluminense. Instituto de Matemática e Estatística. IV. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Resumo

A violência e a criminalidade causam danos e prejuízos à sociedade e ao Estado. O município do Rio de Janeiro, além de sofrer com problemas históricos na área de segurança pública, é afetado por um ciclo que envolve a menor participação na economia do país, indicadores sociais em queda, aumento do desemprego e o consequente aumento da criminalidade, em especial dos roubos. A coleta, estudo e análise de dados de roubos é importante para conseguir entender seus padrões. Pesquisar sobre a diferença de áreas geográficas e outros fatores que possam influenciar a ocorrência do roubo auxilia a prevenção e combate dos mesmos. Nesse contexto, este trabalho teve o objetivo de identificar, compreender e descrever o padrão de comportamento de dados de roubos na cidade do Rio de Janeiro no ano de 2019, através de modelos aditivos generalizados. Nesses modelos, foram incorporados indicadores sociais, econômicos, demográficos, educacionais e geográficos bem como componentes espaciais capazes de representar o padrão espacial dos roubos. Desejava-se avaliar os indicadores e associá-los com a ocorrência de roubos na cidade do Rio de Janeiro, no ano de 2019, utilizando modelos aditivos generalizados para estimar as probabilidades de ocorrência de roubos na cidade do Rio ao nível de quadrículas de 200 x 200 metros e incorporar os efeitos espaciais aos modelos aditivos generalizados a partir de suavizadores na localização geográfica de cada quadrícula. Dados foram obtidos do Instituto de Segurança Pública do Estado do Rio de Janeiro, do Censo demográfico de 2010 do Instituto Brasileiro de Geografia e Estatística, do Instituto Pereira Passos e do OpenStreetMap. As variáveis do modelo final deste trabalho foram selecionadas através de análise descritiva e testes no modelo proposto. Os resultados mostraram que as distâncias até a via rápida mais próxima, escola mais próxima e estação de trem mais próxima, as proporções de domicílios sem moradores do sexo masculino, domicílios em locais onde existe calçada, pessoas de mais de 59 anos de idade, a população média por célula e as coordenadas de longitude e latitude dos centroides das quadrículas possuem efeitos significativos na probabilidade de ocorrência de roubos no município. Dessa forma, a incorporação das informações espaciais ao modelo aditivo generalizado proposto pelo estudo se mostrou importante.

Palavras-chave: Modelos Aditivos Generalizados. *Splines*. Roubos. Rio de Janeiro.

Dedicatória

Dedico este trabalho à minha mãe Ana Suzy, meu pai Fernando, meu irmão Gabriel e minha avó Maria de Lourdes, que formam, hoje, a fundação mais forte minha vida. Dedico, também, aos meus avós Agostino Pigozzo e Vera Hilda, que certamente estão sorrindo em um outo plano.

Agradecimentos

Gostaria de demonstrar minha gratidão à minha família, principalmente à minha mãe, meu pai e meu irmão, por todo suporte, carinho, atenção e incentivos dedicados a mim nessa jornada.

Gostaria de demonstrar minha gratidão, também, às minhas amigas e meus amigos, que com muito apoio, incentivo, paciência e "empurrões"me ajudaram a chegar até aqui. Em especial às minhas nenéns e ao responsável pela inscrição surpresa no ENEM que deu início a esse belo (longo) caminho.

Gostaria de agradecer a todos os professores que participaram da minha formação, em especial aos meus orientadores, Rafael e Jony e minha coordenadora, Ana Maria.

Muito obrigado, de coração, a todos vocês!

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Este capítulo apresenta a motivação para a realização deste trabalho, uma breve revisão de literatura ao tema discutido e os objetivos geral e específico.

1.1 Motivação

O bem-estar dos cidadãos é bastante impactado pela questão da criminalidade. Patrimônios são perdidos e a integridade física das pessoas é colocada em risco. Tal questão torna-se, assim, fator de preocupação tanto aos governantes e autoridades quanto à população de um modo geral. Dessa forma, o crime acaba gerando inúmeros prejuízos à sociedade. Cerqueira et al. (2019) mencionam que, em 2016, a criminalidade causou uma perda de 373 bilhões de reais para o Brasil, o equivalente à $5,9\%$ do PIB nacional. Ademais, novos investimentos são afugentados, investimentos existentes são extintos e a atividade turística é deteriorada, tudo por conta da criminalidade. Estes fatores ocorrem, principalmente, nas grandes cidades do país.

O estado do Rio de Janeiro, por exemplo, vem sofrendo com um ciclo que associa menor participação na economia brasileira, indicadores sociais em queda, violência e desigualdade territorial. Há, também, os efeitos da crise política e econômica dos últimos anos, culminando em grande perda de empregos com carteira assinada, principalmente, no município do Rio de Janeiro. O quadro de violência é, então, agravado. Como resultado: há um aumento da violência numa localidade onde já há problemas históricos e particulares na área de segurança pública (OSORIO; VERSIANI; VEIGA, 2018).

Entretanto, sabe-se que uma boa gestão a respeito da segurança pública pode e deve ser baseada num planejamento operacional que se inicia na sistematização dos dados (BEATO, 2000). Em muitos estados e cidades do Brasil, a análise da criminalidade vem sendo aprimorada. No Rio de Janeiro, por exemplo, o Instituto de Segurança Pública do Estado do Rio de Janeiro iniciou o processo de estruturação, representação e análise

dos dados criminais a partir de geotecnologias. Isso possibilitou identificar áreas de maior risco para diversos crimes, auxiliando a atuação da polícia de forma rápida, inteligente e com custo reduzido (GONÇALVES, 2021).

Diante dos fatos citados, é possível notar a importância de se coletar, estudar e analisar dados referentes a crimes para compreender seus padrões no espaço e tempo. Pesquisar sobre diferença de áreas geográficas e outros fatores que possam influenciar a incidência de determinado acontecimento criminoso surge como um norte para que ações de prevenção e combate dessas práticas sejam melhores elaboradas pelas autoridades competentes.

Nesse sentido, a análise de dados espaciais através de modelos estatísticos que capturem a dependência espacial torna-se fundamental na previsão e identificação de fatores que consigam identificar a dinâmica dos crimes no espaço e no tempo. Esses métodos podem ser utilizados para auxiliar a tomada de decisão dos gestores e criação de políticas públicas.

Em estudos dessa natureza, normalmente, a região de interesse possui subdivisões. Um caminho natural seria fazer análises utilizando modelos estatísticos que capturem a dependência espacial dos dados a partir de componentes estruturadas. Isso, porém, se torna inviável quando a resolução espacial é demasiadamente grande e gera uma dimensão de informações com a qual não é possível trabalhar. Como a incorporação das características espaciais dos dados deste trabalho deve ser considerada na análise, uma possibilidade de resolução dessa questão é a utilização de modelos aditivos generalizados (GAM) que permite a suavização da geolocalização dos dados como variáveis nos modelos. Apesar de não usar componentes espacialmente estruturadas, esses modelos com suavizadores na componente espacial são simples de se estimar e sua utilização em dados de grande dimensão se torna viável.

1.2 Revisão Bibliográfica

Como mencionado anteriormente, a análise dos dados de criminalidade é fundamental para tomada de decisão e geração de políticas de segurança. Na literatura, há diversos trabalhos que abordam o uso de modelos estatísticos para descrição dos padrões desse tipo de dado. Por exemplo, Wang e Brown (2012) utilizam modelos aditivos generalizados espaço-temporais e modelos aditivos generalizados espaço-temporais locais na tentativa de estudar fatores relacionados a crimes e prever incidentes futuros. Os dados avaliados se referem a incidentes de invasões de domicílios em Charlottesville nos EUA entre abril de

2001 e fevereiro de 2005. Além das informações dos incidentes, são incluídas informações geográficas da cidade, como localizações de escolas, rodovias, comércio, etc. e informações demográficas, como população e valores dos imóveis, por exemplo, medidas por setores censitários. Para modelar os incidentes criminais, a cidade foi particionada em quadrículas de 32×32 metros, totalizando 23.089 quadrículas. O intervalo de tempo foi de um mês, totalizando 46 meses. Dessa forma, havia 1.062.094 registros e cada um possuía uma variável resposta indicando se pelo menos um incidente aconteceu na quadrícula naquele mês. O objetivo das modelagens foi prever os locais e momentos dos incidentes futuros e a probabilidade de um incidente ocorrer em um local e mês definidos. Os resultados mostraram que esses modelos propostos pelos autores foram significativamente melhores do que o modelo linear generalizado espacial. Além disso, Wang e Brown (2012) observaram que os efeitos das covariáveis não são lineares.

Em outro estudo, sobre roubos a residências em Amsterdam, Mahfoud, Bhulai e Mei (2018) generalizaram o modelo proposto por Wang e Brown (2012) permitindo interações espaço-temporais mais complicadas através de um modelo aditivo generalizado para modelar a distribuição de probabilidade citada. Um distrito de Amsterdam foi dividido em quadrículas de 125 x 125 metros, e foram levadas em conta apenas as quadrículas correspondentes às áreas urbanas. À cada quadrícula, foram associadas covariáveis correspondentes aos fatores demográficos e socioeconômicos de suas vizinhanças e covariáveis correspondentes à informações geográficas da cidade, como distâncias até as vias expressas, por exemplo. Os autores constataram que uma pequena parte da variação dos dados foi capturada pelo modelo e que esse poder reduzido pode ter sido causado pela alta-resolução dos dados usados. Outra constatação importante foi a que a relação entre as covariáveis e a resposta de seu modelo não eram lineares.

Ainda no contexto de análise de crimes, uma pesquisa nacional, mais recente, investigou homicídios de pessoas negras no Brasil utilizando um modelo aditivo generalizado com uma componente geoespacial. Lizzi et al. (2021) investigaram as razões sociais para o risco de homicídio da população negra nas 27 unidades da federação. Foram utilizados o índice de desenvolvimento humano e seus componentes, índices de desigualdade sociais, uma componente temporal e a componente espacial para estimar o risco de homicídio dentro da população negra no Brasil. A escala de tempo foi de anos, divididos entre 2000 e 2016 e as coordenadas de latitude e longitude dos centroides de cada estado foram usadas como componente espacial. O modelo assumiu que os homicídios possuíam distribuição de Poisson e os resultados mostraram que os homicídios seguiam uma crescente linear durante todo o período estudado e que onde havia menor taxa de longevidade e menor

renda, havia maior risco de homicídio. Lugares caracterizados por maiores desigualdades possuíam, também, maior risco de homicídio. Não foi possível distinguir um padrão nos homicídios da população negra em relação à educação, que poderia, segundo os autores, ser devido à grande disparidade de níveis de educação entre os estados, podendo ser melhorado com informações agregadas em áreas menores, como por cidades ou mesorregiões, ao invés de estados.

1.3 Objetivos

Este trabalho tem o objetivo principal de identificar, compreender e descrever o padrão de comportamento dos dados de roubo[1] na cidade do Rio de Janeiro, através de modelos aditivos generalizados. Nesses modelos, serão incorporados indicadores sociais, econômicos, demográficos e educacionais bem como componentes espaciais capazes de representar o padrão espacial dos roubos.

Surgem como objetivos secundários:

- Avaliar indicadores econômicos, sociodemográficos, educacionais, de saúde e geográficos, e associá-los com a ocorrência de roubos na cidade do Rio de Janeiro;
- Utilizar modelos aditivos generalizados para estimar as probabilidades de ocorrência de roubos na cidade do Rio de Janeiro ao nível de quadrícula de 200×200 metros;
- Incorporar os efeitos espaciais nos modelos aditivos generalizados a partir da aplicação de suavizadores na localização geográfica de cada quadrícula.

1.4 Organização

A organização deste trabalho é dada em cinco capítulos. Neste capítulo inicial, conceitos e contexto são discutidos, uma breve revisão bibliográfica é realizada e os objetivos são detalhados. No segundo capítulo, expõe-se informações sobre os dados que foram reunidos para as análises e as técnicas estatísticas utilizadas nas mesmas. Os resultados obtidos através da metodologia utilizada neste trabalho são discutidos e analisados no terceiro capítulo. No quarto, descreve-se as principais conclusões e considerações oriundas deste estudo. E, por fim, faz-se as referências bibliográficas deste trabalho.

[1] O Instituto de Segurança Pública do Estado do Rio de Janeiro define que roubo, em sua base de dados, representa o agrupamento dos crimes "Patrimônio - Violento - Móvel", "Patrimônio - Violento - Rua" e "Patrimônio - Violento - Fixo".

2 Materiais e Métodos

Neste capítulo, serão descritos os materiais e a metodologia estatística utilizada neste trabalho. Inicialmente, apresentam-se a área de estudo, as fontes de dados e as variáveis estudadas. Em seguida, definem-se os modelos lineares generalizados e suas propriedades bem como o modelo aditivo generalizado e as características da suavização.

2.1 Área de Estudo

A área de estudo considerada para a realização deste trabalho é o município do Rio de Janeiro. Capital do estado, o município do Rio possui uma área territorial de 1.200 km^2, é dividido, atualmente, em 164 bairros e sua população é de 6.775.561 habitantes, de acordo com a estimativa do Instituto Brasileiro de Geografia e Estatística (IBGE) para julho de 2021. A Figura 1 apresenta o município do Rio de Janeiro e seus limites de bairros atuais.

A unidade espacial utilizada neste trabalho é definida por uma malha de quadrículas de 200×200 metros que cobre toda a extensão do município, particionando-o em 31.508 quadrículas no total. Cada uma dessas quadrículas, então, tem suas características particulares que são indicadas pelas variáveis a serem utilizadas no estudo. A Figura 2 ilustra essa partição por quadrículas através de um pequeno recorte da zona sul do município, contendo os bairros de Ipanema e Copacabana.

2.2 Dados

Um dos objetivos deste trabalho é avaliar a associação entre indicadores sociais, demográficos, econômicos, educacionais e geográficos e a probabilidade de ocorrência de roubos. A escolha desses indicadores para comporem as covariáveis do modelo proposto no trabalho se deu através de informações adquiridas pelas experiências e resultados obtidos nos modelos dos trabalhos citados na Seção 1.2 bem como através de sugestões em

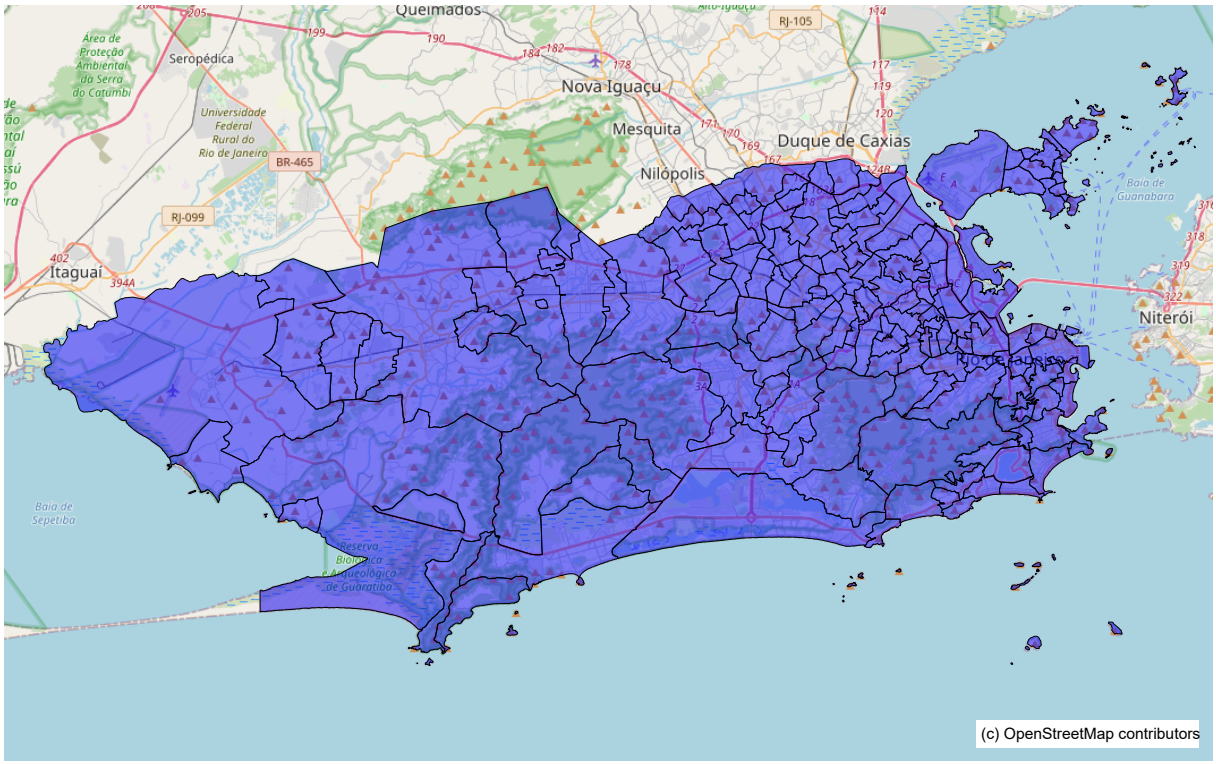


Figura 1: Mapa do município do Rio de Janeiro por bairros.

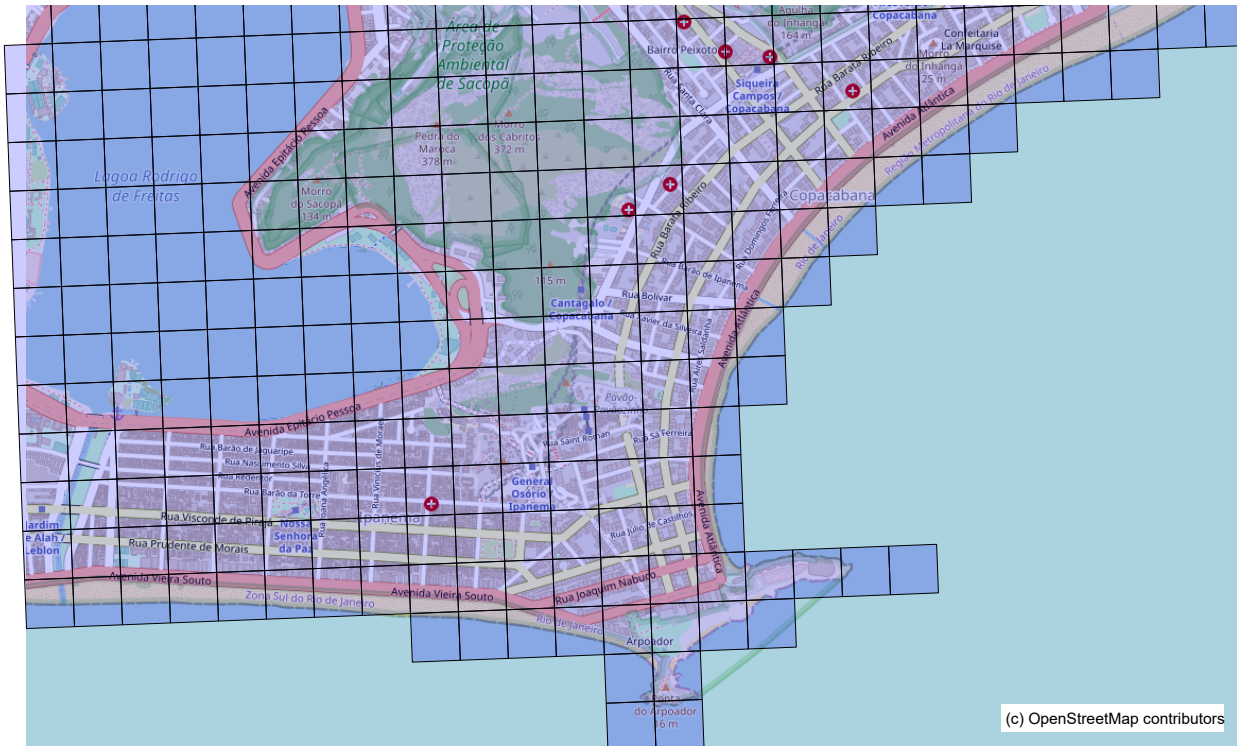


Figura 2: Recorte dos bairros de Ipanema e Copacabana particionados por quadrículas de 200×200 metros.

conversas com equipe de especialistas do Instituto de Segurança Pública do Estado do Rio de Janeiro visando entender a influência que a alocação de esforços governamentais em certas áreas, como policiamento de locais ou infraestrutura e urbanização, poderia ter sobre a probabilidade de ocorrência dos roubos na cidade.

Dessa forma, os dados utilizados aqui foram obtidos de quatro fontes. No que se refere às informações sobre a criminalidade no município do Rio de Janeiro, o Instituto de Segurança Pública do Estado do Rio de Janeiro (ISP-RJ) disponibilizou dados sobre uma série de crimes cometidos no período de 2016 a 2020. As informações sociais, econômicas, demográficas e educacionais, foram obtidas a partir do Censo Demográfico de 2010 disponibilizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Em relação aos dados sobre regiões de favelas, escolas, unidades de saúde e estações de trem, os dados foram obtidos no portal DATA.RIO do Instituto Pereira Passos (IPP). Dados sobre as localizações das vias da cidade do Rio de Janeiro foram obtidos através do *OpenStreetMaps*.

2.2.1 Base de Dados do Instituto de Segurança Pública

Os dados do ISP-RJ estão disponíveis ao nível de quadrículas de 200 x 200 metros, que somadas dão a extensão total do município. Cada entrada do banco de dados do ISP representa uma quadrícula e possui a informação georreferenciada do seu polígono. A partir dos polígonos, é possível obter o centroide de cada quadrícula, na forma de pontos georreferenciados. As coordenadas desses pontos são separadas e dão origem às informações de longitude e latitude para cada uma das quadrículas. As variáveis disponíveis desta fonte são:

- quantidade de roubos na quadrícula i de 200×200 metros, $i = 1, \ldots, 31.508$;
- mês e ano de ocorrência do roubo;
- longitude e latitude da quadrícula i de 200×200 metros, $i = 1, \ldots, 31.508$.

2.2.2 Dados Sociodemográficos e Geográficos

2.2.2.1 Dados do Censo Demográfico

O IBGE realiza decenalmente o Censo Demográfico e sua edição mais recente foi a do ano de 2010. Neste trabalho, serão utilizados dados dos resultados do universo, que

compreendem características dos domicílios e das pessoas investigadas no censo para a totalidade da população. Esses dados estão agregados por setores censitários, que são a menor forma de agregação espacial disponível. O município do Rio de Janeiro possui dados para 10.233 setores censitários.

Os setores censitários são maiores do que as quadrículas de 200×200 metros, escala espacial utilizada neste trabalho, e possuem formas irregulares de acordo com cada localidade. Em decorrência deste fato, há quadrículas que estão localizadas nos limites dos setores, participando de mais de um deles. Fica definido, então, que cada uma das 31.508 quadrículas recebe os valores de seus indicadores de acordo com os valores dos indicadores dos setores censitários aos quais os seus centroides fazem parte.

Os 18 indicadores construídos, para cada setor censitário, a partir da base de dados do Censo Demográfico de 2010 são:

- D101: proporção de domicílios particulares permanentes com abastecimento de água da rede geral

$$D101 = \frac{\text{domicílios particulares permanentes com abastecimento de água da rede geral}}{\text{domicílios particulares permanentes}}$$

- D102: proporção de domicílios particulares permanentes com banheiro de uso exclusivo dos moradores ou sanitário

$$D102 = \frac{\text{domicílios particulares permanentes com banheiro de uso exclusivo dos moradores ou sanitário}}{\text{domicílios particulares permanentes}}$$

- D103: proporção de domicílios particulares permanentes com banheiro de uso exclusivo dos moradores ou sanitário e esgotamento adequado via rede geral de esgoto ou pluvial

$$D103 = \frac{\text{esgotamento sanitário via rede geral de esgoto ou pluvial}}{\text{domicílios particulares permanentes}}$$

- D104: proporção de domicílios particulares permanentes com lixo coletado

$$D104 = \frac{\text{domicílios particulares permanentes com lixo coletado}}{\text{domicílios particulares permanentes}}$$

- D105: proporção de domicílios particulares permanentes com energia elétrica

$$D105 = \frac{\text{domicílios particulares permanentes com energia elétrica}}{\text{domicílios particulares permanentes}}$$

- D106: proporção de domicílios particulares permanentes sem morador do sexo masculino

$$D106 = \frac{\text{domicílios particulares permanentes sem morador do sexo masculino}}{\text{domicílios particulares permanentes}}$$

- E101: proporção de domicílios particulares permanentes onde existe iluminação pública

$$E101 = \frac{\text{próprios onde existe iluminação+alugados onde existe iluminação+cedidos onde existe iluminação}}{\text{domicílios particulares permanentes}}$$

- E102: proporção de domicílios particulares permanentes onde existe pavimentação

$$E102 = \frac{\text{próprios onde existe pavimentação+alugados onde existe pavimentação+cedidos onde existe pavimentação}}{\text{domicílios particulares permanentes}}$$

- E103: proporção de domicílios particulares permanentes onde existe calçada

$$E103 = \frac{\text{próprios onde existe calçada+alugados onde existe calçada+cedidos onde existe calçada}}{\text{domicílios particulares permanentes}}$$

- E104: proporção de domicílios particulares permanentes onde existe arborização

$$E104 = \frac{\text{próprios onde existe arborização+alugados onde existe arborização+cedidos onde existe arborização}}{\text{domicílios particulares permanentes}}$$

- E105: proporção de domicílios particulares permanentes onde existe esgoto a céu aberto

$$E105 = \frac{\text{próprios existe esgoto a céu aberto+alugados existe esgoto a céu aberto+cedidos existe esgoto a céu aberto}}{\text{domicílios particulares permanentes}}$$

- E106: proporção de domicílios particulares permanentes onde existe lixo acumulado nos logradouros

$$E106 = \frac{\text{próprios onde existe lixo acumulado+alugados onde existe lixo acumulado+cedidos onde existe lixo acumulado}}{\text{domicílios particulares permanentes}}$$

- P301: proporção de pessoas residentes de cor ou raça branca

$$P301 = \frac{\text{pessoas residentes de cor ou raça branca}}{\text{pessoas residentes}}$$

- P1301: proporção de pessoas com menos de 18 anos de idade

$$P1301 = \frac{\text{total de pessoas com menos de 18 anos de idade}}{\text{pessoas residentes}}$$

- P1302: proporção de pessoas com mais de 59 anos de idade

$$P1302 = \frac{\text{total de pessoas com mais de 59 anos de idade}}{\text{pessoas residentes}}$$

- DR01: proporção de domicílios particulares com rendimento nominal mensal domiciliar per capita de até 1 salário mínimo

$$DR01 = \frac{\text{domicílios particulares com rendimento nominal mensal domiciliar per capita de até 1 salário}}{\text{domicílios particulares permanentes+domicílios particulares improvisados}}$$

- DR02: proporção de domicílios particulares com rendimento nominal mensal domiciliar per capita de mais de 10 salários mínimos

$$DR02 = \frac{\text{domicílios particulares com rendimento nominal mensal domiciliar per capita de mais de 10 salários}}{\text{domicílios particulares permanentes+domicílios particulares improvisados}}$$

- PQ1: população média por quadrícula

$$PQ1 = \frac{\text{pessoas residentes}}{\text{número de quadrículas}}$$

2.2.2.2 Dados do Instituto Pereira Passos

Dados de Áreas de Favelas

A prefeitura da cidade do Rio de Janeiro, através do Instituto Pereira Passos, mantém o portal DATA.RIO com diversas informações, publicações e dados sobre o município do Rio. Desse portal é possível obter a base de dados com as localizações e limites das favelas do município, com um total de 1074 favelas. Tais localizações e limites são obtidos através do georreferenciamento dos polígonos de cada favela. A partir desses polígonos, há um cruzamento com os pontos georreferenciados dos centroides das quadrículas para que sejam calculadas as distâncias dos pontos dos centroides de cada quadrícula até o limite do polígono de cada favela. Nesse sentido, define-se a seguinte variável:

- dist_fav: a menor distância do ponto do centroide da quadrícula i até os limites dos polígonos das favelas.

As geolocalizações dos polígonos das favelas são apresentadas na Figura 3.

Dados sobre Educação

O IPP disponibiliza, também, bancos de dados com as localizações das escolas municipais, estaduais e federais situadas no município do Rio de Janeiro. São 1.540 escolas municipais, 364 escolas estaduais e 27 escolas federais. As localizações são obtidas através

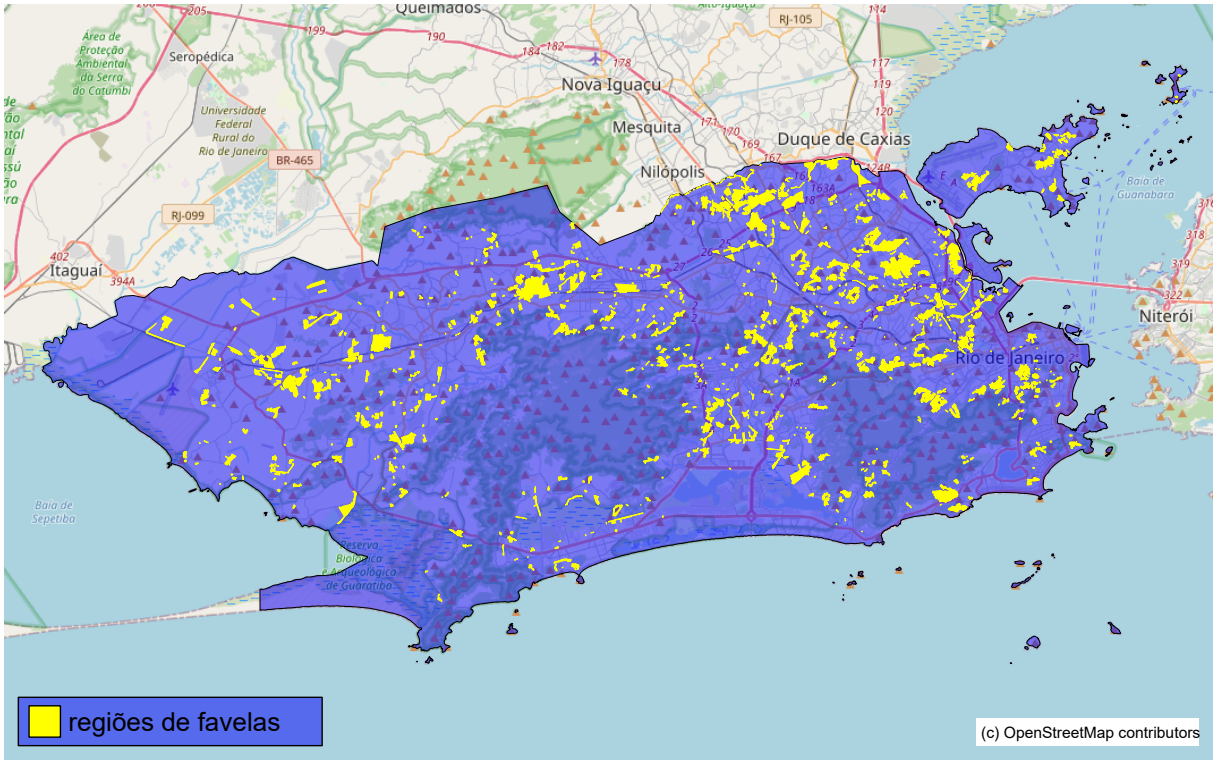


Figura 3: Localização das favelas no mapa do município do Rio de Janeiro.

do georreferenciamento dos pontos de cada escola. A partir desses pontos, há um cruzamento com os pontos georreferenciados dos centroides das quadrículas para que sejam calculadas as distâncias dos pontos dos centroides de cada quadrícula até os pontos georreferenciados de cada escola. Dessa forma, define-se a seguinte variável:

- dist_esc: menor distância do ponto do centroide da quadrícula i até os pontos georreferenciando escolas.

As localizações das escolas são apresentadas na Figura 4.

Dados sobre Saúde

Do mesmo modo, é possível encontrar bases de dados com as localizações de unidades de saúde municipais, estaduais e federais disponibilizadas no portal do IPP. São 688 unidades municipais, 93 unidades estaduais e 47 federais. As localizações são obtidas através do georreferenciamento dos pontos de cada unidade de saúde. A partir desses pontos, há um cruzamento com os pontos georreferenciados dos centroides das quadrículas para que sejam calculadas as distâncias dos pontos dos centroides de cada quadrícula até os pontos georreferenciados de cada unidade de saúde. Assim, define-se a seguinte variável:

- dist_us: menor distância do ponto do centroide da quadrícula i até os pontos geor-

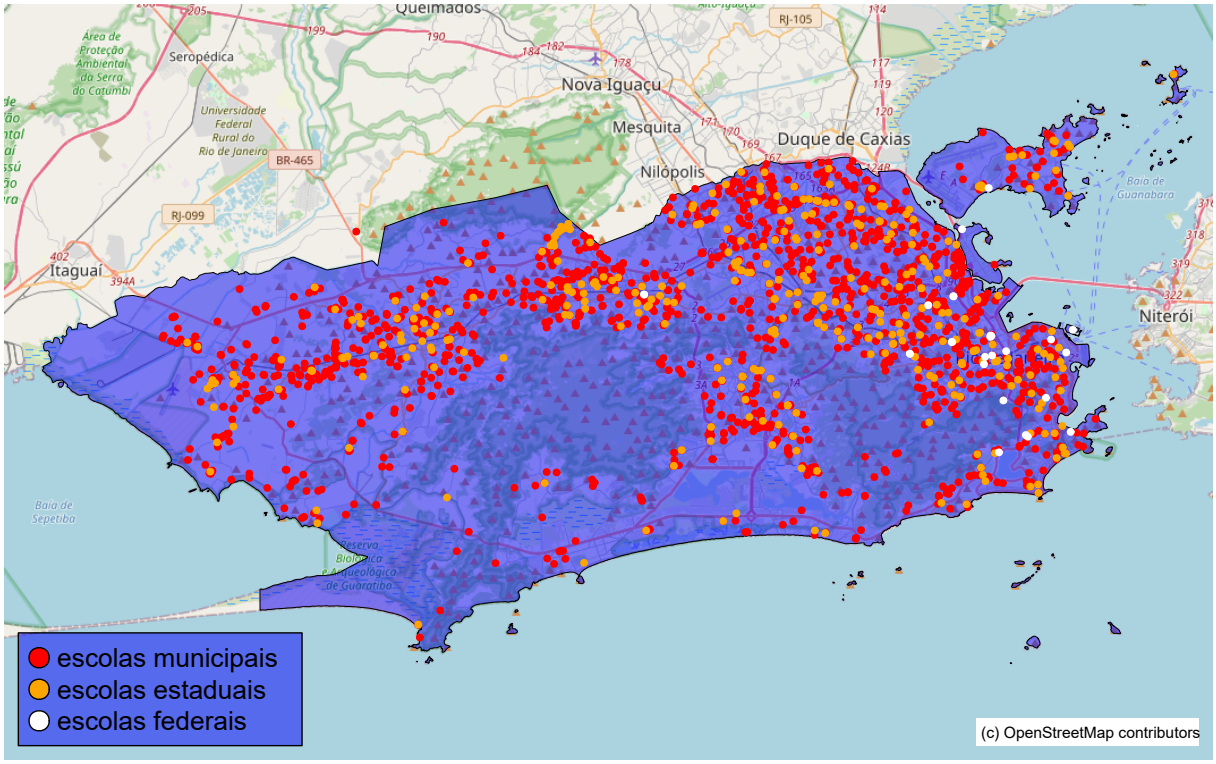


Figura 4: Localização das escolas no mapa do município do Rio de Janeiro.

referenciando unidades de saúde.

As localizações das unidades de saúde são apresentadas na Figura 5.

Dados das Estações de Trem

Mais uma informação que encontra-se no portal do IPP é a base de dados com as localizações das estações de trem de passageiros do município do Rio. Existem 60 estações dentro dos limites da cidade do Rio de Janeiro. As localizações são obtidas através do georreferenciamento dos pontos de cada estação de trem. A partir desses pontos, há um cruzamento com os pontos georreferenciados dos centroides das quadrículas para que sejam calculadas as distâncias dos pontos dos centroides de cada quadrícula até os pontos georreferenciados de cada estação de trem. Define-se, dessa forma, a seguinte variável:

- dist_trem: menor distância do ponto do centroide da quadrícula i até os pontos georreferenciando estações de trem.

As localizações das estações de trem de passageiros são apresentadas na Figura 6.

Dados das Estações do BRT

A última informação utilizada do portal do IPP é a base de dados com as localizações

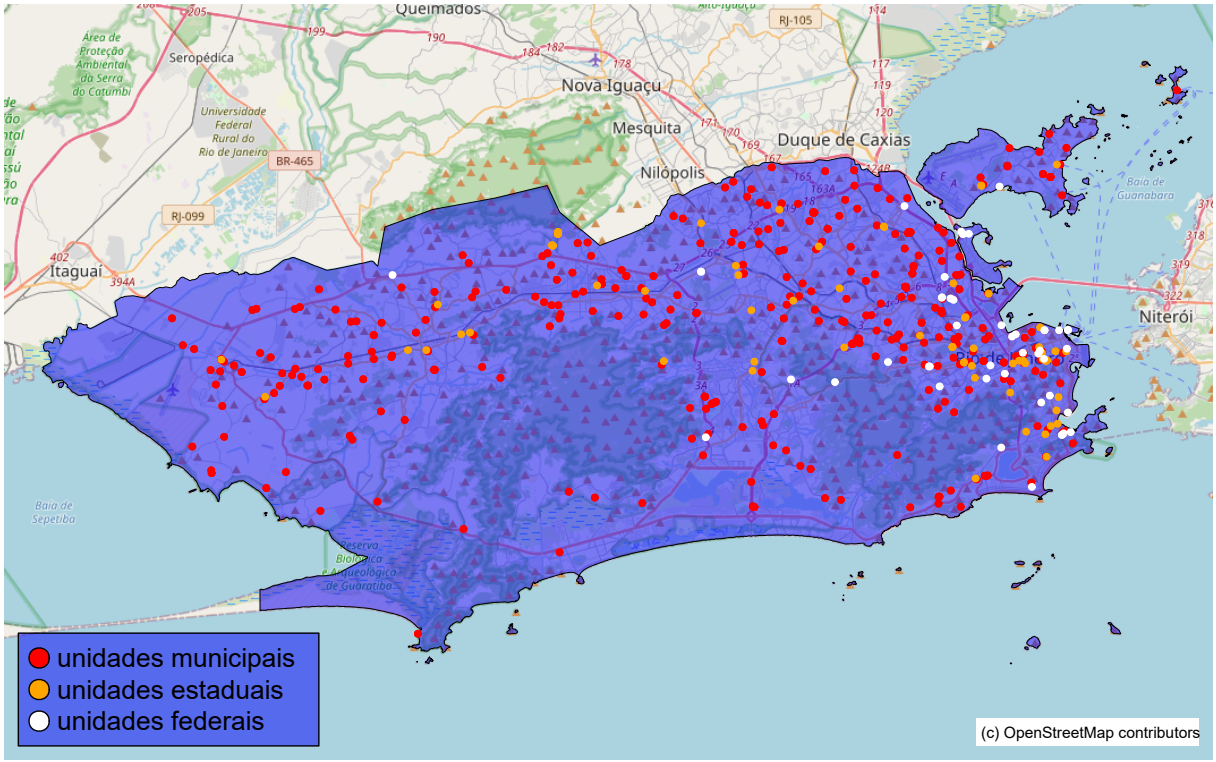


Figura 5: Localização das unidades de saúde no mapa do município do Rio de Janeiro.

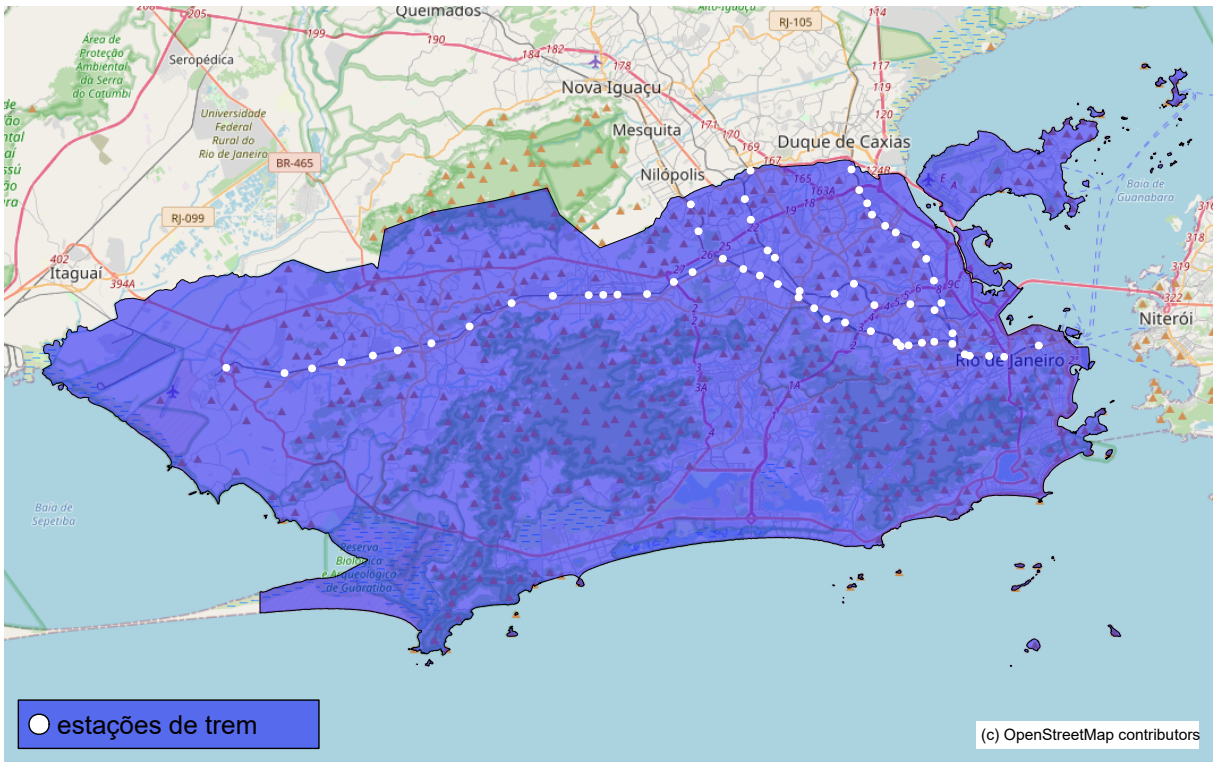


Figura 6: Localização das estações de trem de passageiros no mapa do município do Rio de Janeiro.

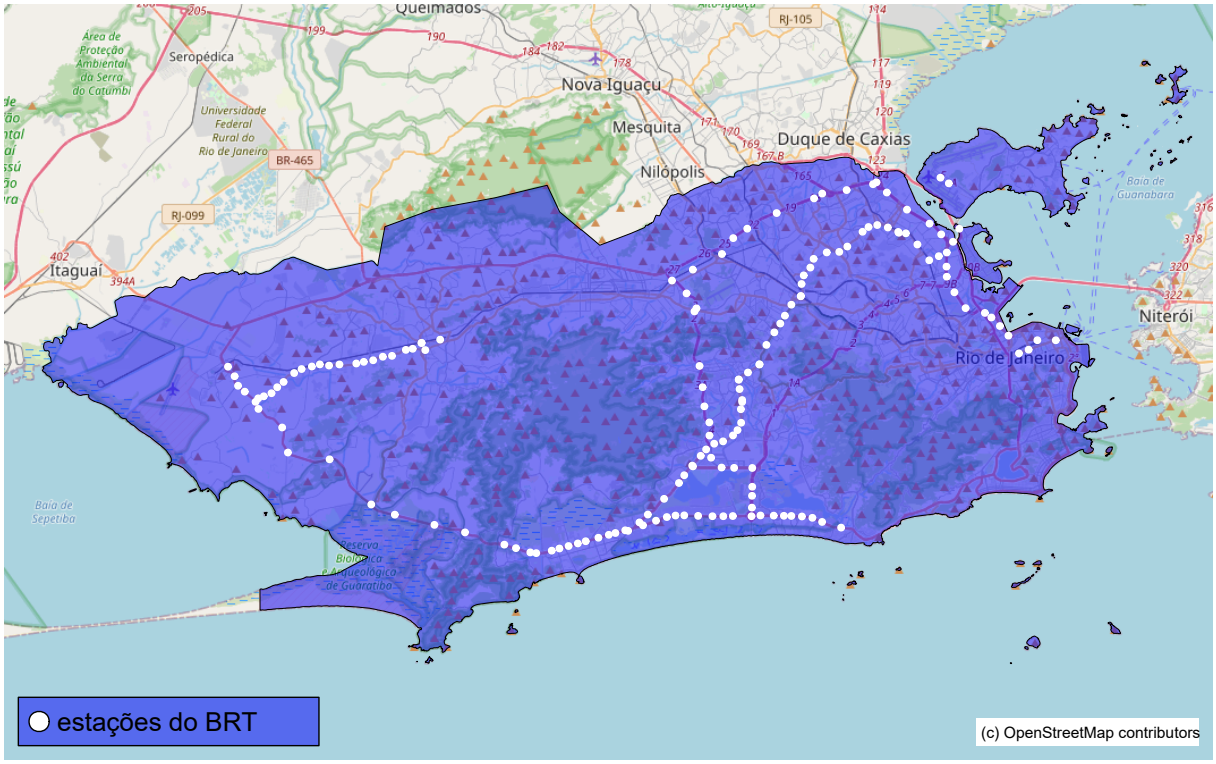


Figura 7: Localização das estações do BRT no mapa do município do Rio de Janeiro.

das estações do BRT do município do Rio. Existem 161 estações dentro dos limites do município. As localizações são obtidas através do georreferenciamento dos pontos de cada estação do BRT. A partir desses pontos, há um cruzamento com os pontos georreferenciados dos centroides das quadrículas para que sejam calculadas as distâncias dos pontos dos centroides de cada quadrícula até os pontos georreferenciados de cada estação do BRT. Assim, a seguinte variável é definida:

- dist_brt: menor distância do ponto do centroide da quadrícula i até os pontos georreferenciando estações do BRT.

As localizações das estações do BRT são apresentadas na Figura 7.

2.2.2.3 Dados do *OpenStreetMap*

O *OpenStreetMap* (OpenStreetMap Foundation, 2022) é uma iniciativa para criar e disponibilizar dados geográficos acessíveis a todos, desenvolvida por uma comunidade voluntária de mapeadores que contribuem e mantêm sua base de dados atualizada. A Fundação *OpenStreetMap* é uma organização internacional sem fins lucrativos que apoia mas não controla o *OpenStreetMap*. Ela é dedicada a encorajar o crescimento, desenvolvimento e distribuição de dados geoespaciais para o uso e compartilhamento de todos.

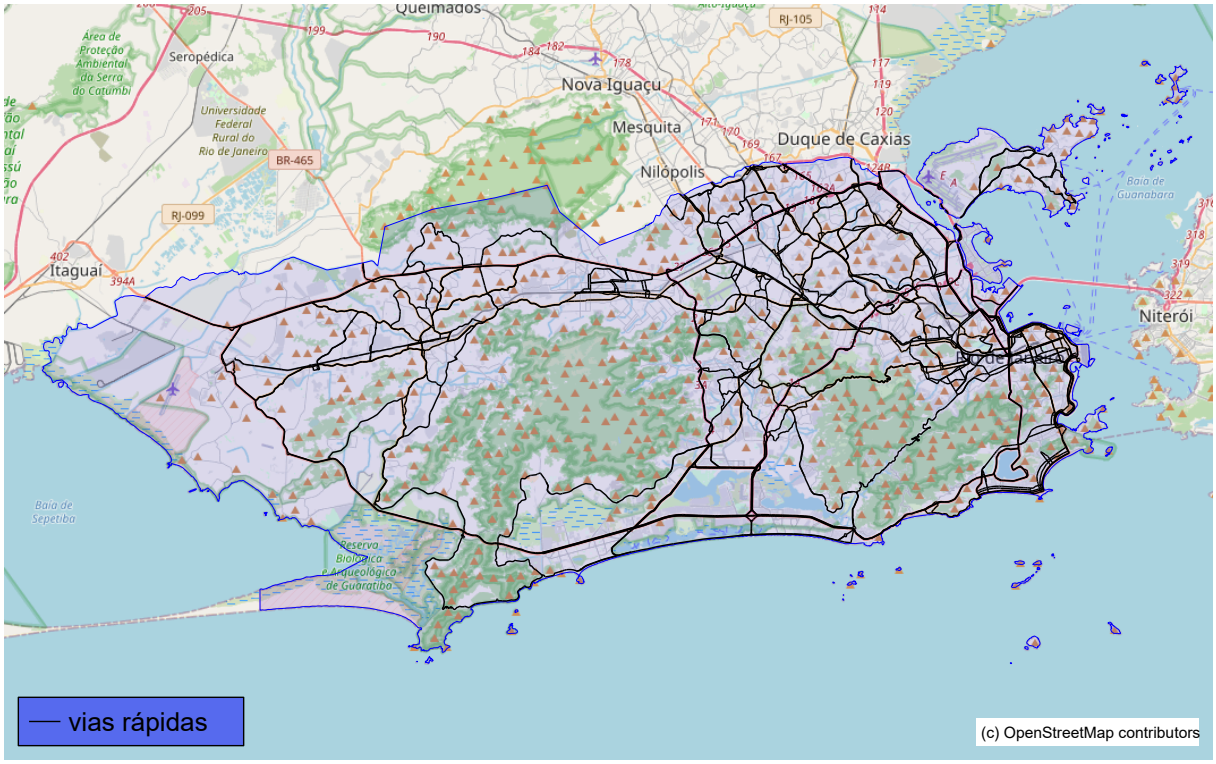


Figura 8: Localização das vias rápidas no mapa do município do Rio de Janeiro.

Através de seu portal é possível baixar o banco de dados que contém informações das localizações de cada via da cidade do Rio. Essas localizações são dadas pelo georreferenciamento das linhas de cada via. Existe, também, a informação do tipo de via. Assim, é possível separar apenas as vias rápidas do município do Rio de Janeiro. Entende-se aqui, por vias rápidas as vias estruturais expressas, como a Linha Amarela, estruturais não-expressas, como a Avenida Brasil, arteriais principais, como a Avenida Presidente Vargas, arteriais secundárias, como a Estrada do Itanhangá, e vias coletoras da cidade, como a Avenida Lúcio Costa. Ou seja, as vias locais, que são as ruas mais simples dentro de cada bairro, não fazem parte desse grupo. A partir das linhas georreferenciadas de cada via rápida, há um cruzamento com os pontos georreferenciados dos centroides das quadrículas para que sejam calculadas as distâncias dos pontos dos centroides de cada quadrícula até as linhas de cada via rápida. Define-se, assim, a seguinte variável:

- dist_via: menor distância entre o centroide da quadrícula i até as linhas georreferenciadas das vias rápidas.

As localizações das vias rápidas do município do Rio são apresentadas na Figura 8.

2.3 Modelos Lineares Generalizados (MLG)

A família de modelos lineares generalizados surge como uma alternativa aos modelos lineares mais usuais, que assumem normalidade dos dados e constância na variância dos mesmos, para tratar de desfechos de naturezas diferentes. Dessa forma, pode-se utilizá-los quando trabalha-se com dados de contagem ou dados binários (que possuem distribuições de probabilidade diferentes da normal), por exemplo. Nos MLG, porém, a distribuição da variável resposta precisa pertencer à família exponencial de distribuições, que engloba distribuições como a normal, binomial e Poisson, entre outras. Além disso, a relação entre a variável resposta e as variáveis explicativas não precisa ser diretamente linear (MONTGOMERY; PECK; VINING, 2012)(DOBSON; BARNETT, 2018)(WOOD, 2017).

Para definição do conceito de família exponencial, considere uma variável aleatória Y, cuja distribuição de probabilidade depende apenas de um parâmetro θ. A distribuição pertence à família exponencial de distribuições se sua função de densidade puder ser escrita da forma:

$$f(y;\theta) = e^{a(y)b(\theta)+c(\theta)+d(y)}. \tag{2.1}$$

A distribuição é escrita na forma canônica quando $a(y) = y$. Chama-se de parâmetro natural da distribuição a função $b(\theta)$. Pode haver a situação em que outros parâmetros existam, neste caso, são tratados como conhecidos e chamados de parâmetros de ruído (DOBSON; BARNETT, 2018).

Para ilustrar a definição de família exponencial, a distribuição Bernoulli é apresentada no Exemplo 2.1.

Exemplo 2.1 *Distribuição Bernoulli*

Seja uma variável aleatória discreta Y, onde $Y \sim Bernoulli(\theta)$. Sua função de probabilidade é escrita da forma:

$$f(y,\theta) = \theta^y(1-\theta)^{1-y},$$

onde y assume os valores 0 ou 1. Essa função pode ser reescrita:

$$f(y,\theta) = e^{ylog(\theta)+(1-y)log(1-\theta)} = e^{ylog(\theta)-ylog(1-\theta)+log(1-\theta)}$$

$$f(y,\theta) = e^{ylog(\frac{\theta}{1-\theta})+log(1-\theta)}, \tag{2.2}$$

onde $a(y) = y$, $b(\theta) = log(\frac{\theta}{1-\theta})$, $c(\theta) = log(1-\theta)$ *e* $d(y) = 0$. *Assim, a distribuição Bernoulli pertence à família exponencial na forma canônica.*

Para definir um modelo linear generalizado, tem-se $Y_1, Y_2, \ldots, Y_N$, todos com uma mesma distribuição de probabilidade petencente à família exponencial na forma canônica, dependente de um único parâmetro θ_i, $i = 1, \ldots, N$, que pode possuir valores diferentes para cada i (DOBSON; BARNETT, 2018).

Três elementos compõem os MLG: um componente aleatório que são as variáveis resposta, $Y_1, \ldots, Y_N$, com mesmo tipo de distribuição; um componente sistemático definindo o preditor linear, η_i,

$$\eta_i = \boldsymbol{X}_i^T \boldsymbol{\beta}, \tag{2.3}$$

onde, $\boldsymbol{X}_i^T$ é a i-ésima linha da matriz dos valores das variáveis explicativas $\boldsymbol{X}$,

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_N^T \end{bmatrix} = \begin{bmatrix} 1 & \boldsymbol{x}_{11} & \cdots & \boldsymbol{x}_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \boldsymbol{x}_{N1} & \cdots & \boldsymbol{x}_{Np} \end{bmatrix} \tag{2.4}$$

e o vetor $\boldsymbol{\beta}$ é um vetor de parâmetros desconhecidos,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}; \tag{2.5}$$

e uma função de ligação, monótona e diferenciável, que relaciona a média da i-ésima observação com seu preditor linear,

$$g(\mu_i) = \eta_i = \boldsymbol{X}_i^T \boldsymbol{\beta}, \tag{2.6}$$

onde $\mu_i = \mathbb{E}(Y_i)$.

2.3.1 Estimação dos parâmetros

Os parâmetros $\boldsymbol{\beta}$ de um MLG podem ser estimados utilizando métodos baseados na máxima verossimilhança. Estes, são métodos numéricos, iterativos e fundamentados no algoritmo de Newton-Raphson. A equação iterativa para obtenção dos estimadores pontuais é:

$$\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \hat{\boldsymbol{\beta}}^{(m)} = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{z}, \tag{2.7}$$

onde $\boldsymbol{X}$ é a matriz desenho apresentada na Equação (2.4), $\hat{\boldsymbol{\beta}}^{(m)}$ é o vetor com as estimativas na m-ésima iteração, $\boldsymbol{W}$ é a matriz diagonal com elementos

$$w_{ii} = \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \tag{2.8}$$

e $\boldsymbol{z}$ tem elementos

$$z_i = \sum_{k=1}^{p} x_{ik} \hat{\beta}_k^{(m-1)} + (y_i - \mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i} \right). \tag{2.9}$$

Para obter as estimativas, estabelece-se um valor inicial para $\hat{\boldsymbol{\beta}}^{(0)}$, utiliza-o em $\boldsymbol{z}$ e $\boldsymbol{W}$ e obtem-se $\hat{\boldsymbol{\beta}}^{(1)}$ através da equação iterativa. A estimativa pontual $\hat{\boldsymbol{\beta}}^{(m)}$ é definida quando a diferença entre $\hat{\boldsymbol{\beta}}^{(m-1)}$ e $\hat{\boldsymbol{\beta}}^{(m)}$ for consideravelmente pequena.

Para maiores informações sobre o método de Newton-Raphson, a equação iterativa e obtenção de seus valores, consultar Dobson e Barnett (2018).

2.4 Modelos Aditivos Generalizados

Os modelos aditivos generalizados, do inglês *generalized additive models (GAM)*, podem ser considerados uma variação dos modelos lineares generalizados de forma a flexibilizar as relações entre a média da variável resposta e as variáveis explicativas usualmente utilizadas num MLG. Nos GAM, o preditor linear η inclui a soma de funções suavizadas das variáveis explicativas. Tais funções são não-paramétricas, não necessariamente conhecidas, mas estimadas a partir dos dados, o que possibilita uma configuração em que os dados guiam a relação que possuem com o preditor.

Dessa forma, pode-se representar o modelo aditivo generalizado como:

$$g(\mu_i) = \boldsymbol{A_i}\boldsymbol{\delta} + f_1(x_{1i}) + f_2(x_{2i}) + ... \tag{2.10}$$

onde $\mu_i = \mathbb{E}(Y_i)$ e Y_i é a variável resposta com distribuição pertencente à família exponencial. $\boldsymbol{A_i}$ é uma linha da matriz do modelo para componentes estritamente paramétricos, $\boldsymbol{\delta}$ é o vetor de parâmetros correspondente e f_j são funções suavizadas das covariáveis x_k.

A flexibilidade e conveniência na especificação da dependência entre a variável resposta e as variáveis explicativas nos GAM demandam duas questões, precisa-se representar as funções suavizadas de alguma forma e decidir o quão suaves serão.

Os modelos aditivos generalizados não estimam parâmetros de um modelo, como nos MLG, eles estimam as funções suavizadas f_j expressas como uma curva para cada

covariável.

2.4.1 Suavizadores

Para fazer a estimação das funções suavizadas f utilizando os mesmos métodos usados na estimação de modelos lineares, é necessário que f seja escrita de forma propícia, como um modelo linear. Assim, deve-se escolher uma base, o que significa escolher o espaço de funções onde f será algum elemento pertencente a ele. A escolha dessa base significa escolher funções base que entram na análise, de fato, e são tratadas como conhecidas. Representando $b_j(x)$ como uma dessas funções base, que será o x transformado dependendo da base escolhida, define-se, então, uma função suavizada qualquer como:

$$f(x) = \sum_{j=1}^{k} b_j(x)\gamma_j, \tag{2.11}$$

onde os γ_j representam os valores dos parâmetros desconhecidos (WOOD, 2017). Na literatura, há diversas bases definidas. No Exemplo 2.2 apresenta-se um simples exemplo da base polinomial.

Exemplo 2.2 *Base polinomial*

Considere um modelo com uma função e uma covariável da forma:

$$y_i = f(x_i) + \epsilon_i,$$

onde y_i é a variável resposta, x_i a covariável, f uma função suavizada e os ϵ_i são variáveis aleatórias independentes, $\epsilon_i \sim N(0, \sigma^2)$.

Suponha, agora, que existam indicações de que f seja um polinômio de grau 3, de forma que o espaço formado por polinômios de grau 3 e menores contenha f. Uma base para esse espaço é $b_1(x) = 1$, $b_2(x) = x$, $b_3(x) = x^2$ e $b_4(x) = x^3$.

Substituindo a base na Equação (2.11)*, tem-se:*

$$f(x) = \gamma_1 + \gamma_2 x + \gamma_3 x^2 + \gamma_4 x^3.$$

Substituindo esta última, na equação do modelo, resulta em:

$$y_i = \gamma_1 + \gamma_2 x_i + \gamma_3 x_i^2 + \gamma_4 x_i^3 + \epsilon_i,$$

que é um modelo simples de se resolver.

A base polinomial exemplificada acima pode ser útil se for apropriada aos dados. Bases polinomiais possuem limitações devido à natureza global de seus ajustes, não sendo tão flexíveis para suavizar curvas. A base polinomial considera o domínio inteiro da função f, e, assim, ela pode se ajustar não tão bem em alguns locais e oscilar demais em locais onde não deveria, não sendo capaz de captar relações mais complexas. Para exemplificar essa limitação, um pequeno exemplo simulado é apresentado em Exemplo 2.3.

Exemplo 2.3 *Diferentes ajustes da regressão polinomial para uma função*

A Figura 9 mostra o gráfico de dispersão com 100 observações de duas variáveis simuladas, x e y, onde y foi obtida partindo da função descrita pela equação

$$f(x) = 5 + e^{-\frac{x}{3}} + 1.9 \cdot e^{-0.1 \cdot (x-3)^2} - 0.9 \cdot e^{-0.8 \cdot (x-6)^2} - 0.3 \cdot e^{-0.5 \cdot (x-12)^2}$$

acrescida de um erro ϵ, com $\epsilon \sim N(0; 0, 15)$. Foram ajustadas quatro curvas utilizando regressão polinomial com graus 3 (cinza), 6 (azul claro), 10 (roxo) e 15 (dourado). A curva real da função se apresenta pela linha pontilhada em preto. Percebe-se que mesmo um polinômio de grau 15 não consegue captar a complexidade da curva original, havendo locais em que há um ajuste relativamente bom, mas outros com ajuste ruim, onde a curva em amarelo oscila de maneira diferente da curva real.

No Exemplo 2.4, é apresentada uma abordagem melhor do que apenas utilizar uma base polinomial para resultar em uma regressão polinomial. Essa abordagem introduz a intuição necessária para o conceito de *splines*, mais especificamente *cubic splines* (*splines* cúbicos), método utilizado neste trabalho para a estimação das funções suavizadas necessárias para compor o modelo proposto.

Exemplo 2.4 *Polinômios cúbicos "por partes"(Piecewise-cubic polynomials)*

A Figura 10 mostra o gráfico de dispersão para dados simulados, agora com 500 observações das variáveis x e y, com y obtida através da equação

$$f(x) = 5 + e^{-\frac{x}{3}} + 1.9 \cdot e^{-0.1 \cdot (x-3)^2} - 0.9 \cdot e^{-0.8 \cdot (x-6)^2} - 0.3 \cdot e^{-0.5 \cdot (x-12)^2}$$

acrescida de um erro ϵ, com $\epsilon \sim N(0; 0, 15)$, da mesma maneira que no Exemplo 2.3.

Para tentar obter um ajuste melhor do que a regressão polinomial, pode-se pensar em separar esse gráfico em partes e ajustar um polinômio de grau 3 em cada uma delas para

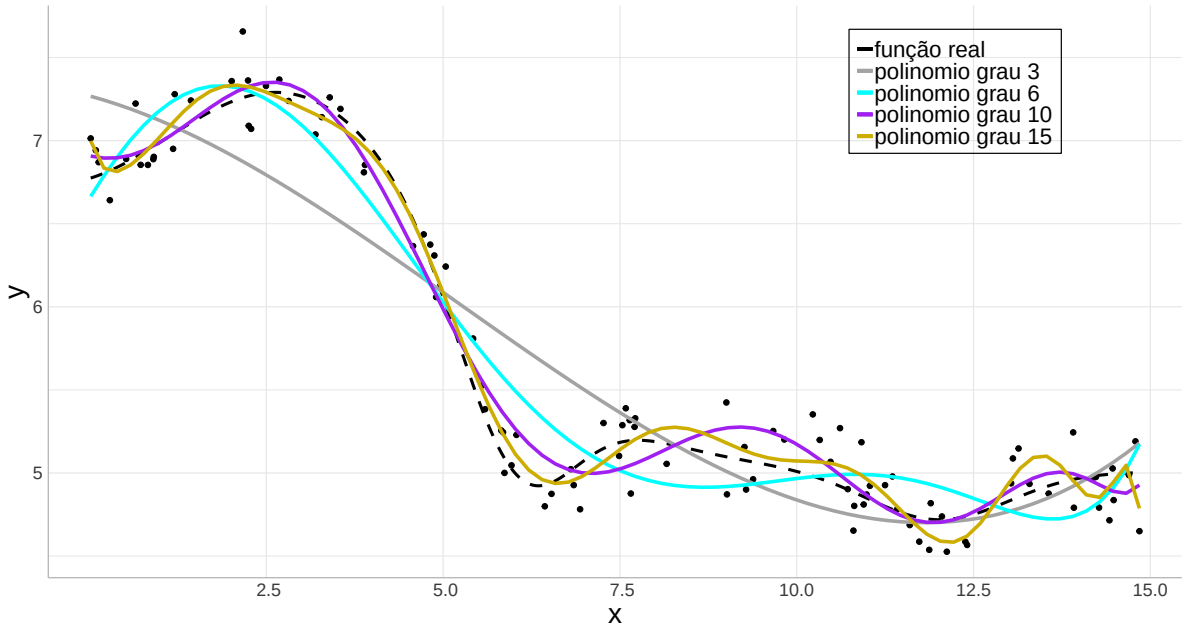


Figura 9: Gráfico de dispersão dos dados simulados para as variáveis x e y com curvas ajustadas por polinômios de diversos graus.

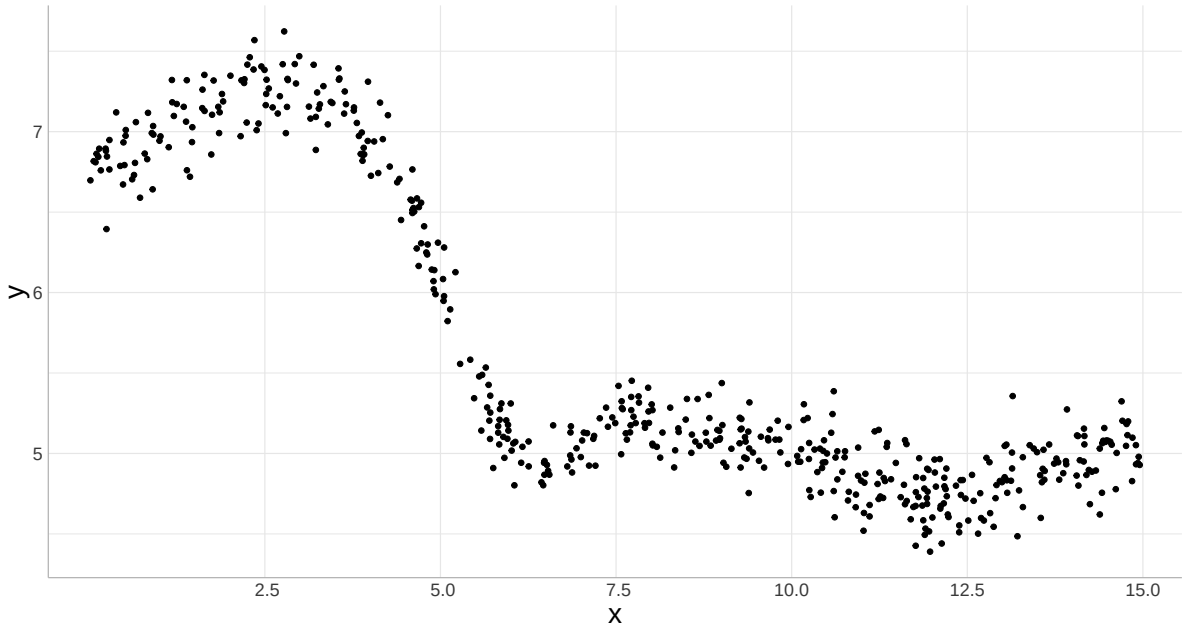


Figura 10: Gráfico de dispersão dos dados para as variáveis x e y.

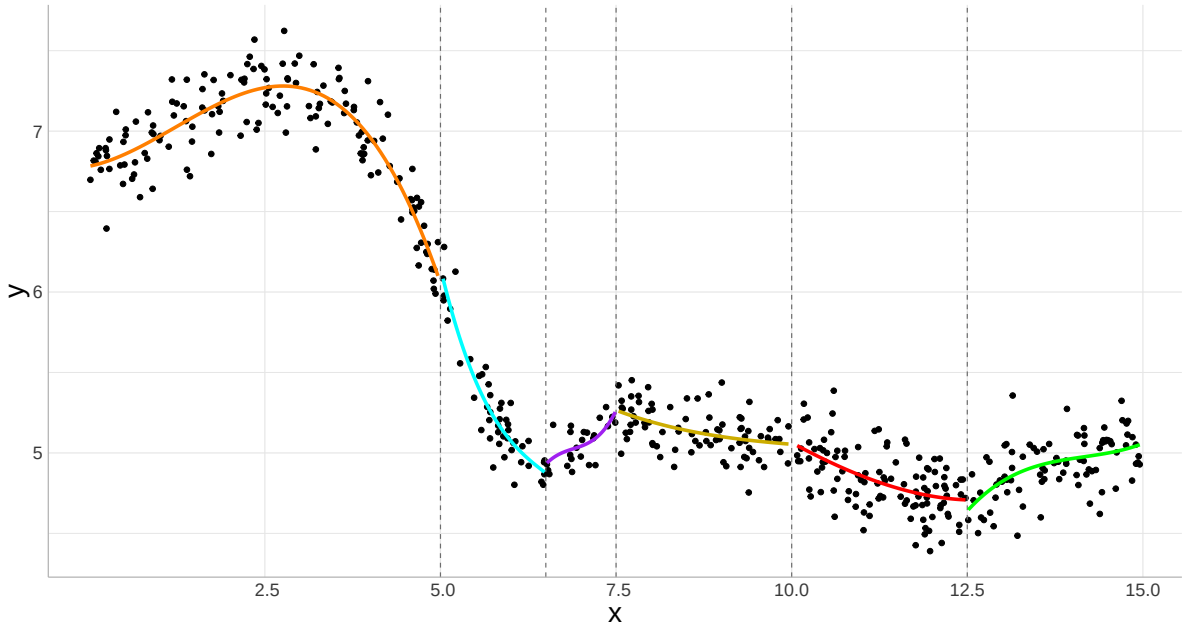


Figura 11: Regressões polinomiais de grau 3 ajustadas dentro de cada intervalo, $[0, 5]$, $(5; 6, 5]$, $(6, 5; 7, 5]$, $(7, 5; 10]$, $(10, 12, 5]$ e $(12, 5; 15]$.

estimar uma curva para a função. São escolhidas, de forma arbitrária, seis partes, que terão regiões definidas por seis intervalos de pontos: $[0, 5]$, $(5; 6, 5]$, $(6, 5; 7, 5]$, $(7, 5; 10]$, $(10; 12, 5]$ *e* $(12, 5; 15]$. *Os limites desses intervalos serão chamados de nós, sendo compostos por cinco nós internos (5, 6,5, 7,5, 10 e 12,5) e dois nós externos (0 e 15). Dessa forma, realiza-se uma regressão polinomial de grau 3 dentro de cada intervalo e obtém-se a Figura 11. É fácil perceber as descontinuidades e os saltos que ocorrem nos nós quando usadas apenas as bases polinomiais simples,* $b_1(x) = 1$, $b_2(x) = x$, $b_3(x) = x^2$ *e* $b_4(x) = x^3$ *para cada um dos ajustes realizados, não tornando essa abordagem muito desejável. Porém, se for utilizada uma base que garanta a continuidade da curva e a não ocorrência de saltos ou mudanças bruscas de curvatura nesses nós, possivelmente, uma boa curva estimada seria obtida. Esses tipos de base são chamadas de splines e, no caso deste exemplo, será usado uma spline cúbica com expressão paramétrica*

$$f(x) = \gamma_1 + \gamma_2 x + \gamma_3 x^2 + \gamma_4 x^3 + \gamma_5 (x - 5)_+^3 + \gamma_6 (x - 6,5)_+^3 + \gamma_7 (x - 7,5)_+^3 + \gamma_8 (x - 10)_+^3 + \gamma_9 (x - 12,5)_+^3,$$

onde

$$(x - c)_+ = \begin{cases} x - c, & x > c \\ 0, & x \leq c \end{cases}$$

e os números 5*,* $6,5$*,* $7,5$*,* 10*, e* $12,5$ *presentes nos cinco últimos termos da soma são os valores dos nós internos. É possível decompor esta expressão como uma combinação linear de 9 funções base* $b_j(x)$*:* $b_1(x) = 1$*,* $b_2(x) = x$*,* $b_3(x) = x^2$*,* $b_4(x) = x^3$*,* $b_5(x) = (x-5)^3_+$*,* $b_6(x) = (x-6,5)^3_+$*,* $b_7(x) = (x-7,5)^3_+$*,* $b_8(x) = (x-10)^3_+$ *e* $b_9(x) = (x-12,5)^3_+$ *podendo, assim, escrever a função na forma da Equação* (2.11)*:*

$$f(x) = \sum_{j=1}^{9} b_j(x)\gamma_j.$$

Agora, é necessário estimar um modelo de regressão linear com equação:

$$y_i = \gamma_1 + \gamma_2 x_i + \gamma_3 x_i^2 + \gamma_4 x_i^3 + \gamma_5 (x_i - 5)^3_+ + \gamma_6 (x_i - 6,5)^3_+ + \gamma_7 (x_i - 7,5)^3_+ +$$
$$\gamma_8 (x_i - 10)^3_+ + \gamma_9 (x_i - 12,5)^3_+ + \epsilon_i$$

para estimar os parâmetros γ_j *e, desta forma, obter uma equação que representa a curva ajustada para os dados. A Figura 12 mostra, em azul claro, a curva ajustada que representa a função suavizada. A curva vermelha pontilhada representa a função real. Percebe-se que se trata de uma boa estimação, a curva se ajusta bem aos dados e captura os movimentos da curva real. Há diferença, porém, se troca-se a posição de um dos nós, por exemplo, de 6,5 para 2,5. A curva pontilhada em roxo, na Figura 12, mostra que esta não possui a mesma qualidade da anterior. Apesar de não ser um ajuste ruim, a curva oscila onde não deveria e nao capta os movimentos da curva real da melhor maneira.*

Apesar do método mostrado no Exemplo 2.4 oferecer uma boa curva estimada, existem problemas. O primeiro referente à quantidade de nós a serem utilizados e suas posições. De fato, quanto mais nós, mais flexibilidade para as funções base ajustarem a curva, mas se houver flexibilidade demais pode ocorrer um sobreajuste e a curva oscilar mais do que deveria. O segundo problema vem com a escolha das funções base que representam os *splines* dados os nós escolhidos. Há o risco de serem acrescentados muitos parâmetros, e, com isso, a matriz do modelo pode apresentar dimensões muito elevantes e gerar problemas práticos e computacionais (HASTIE; TIBSHIRANI, 1990).

Para resolver o problema do sobreajuste com a possível oscilação excessiva da curva e oferecer uma maior eficiência computacional, utiliza-se a técnica de penalizar as oscilações. Essa técnica consiste em definir as funções base com uma dimensão fixa, ligeiramente maior do que se julgaria necessário e controlar a suavização através de uma penalização à estimação por mínimos quadrados (WOOD, 2017).

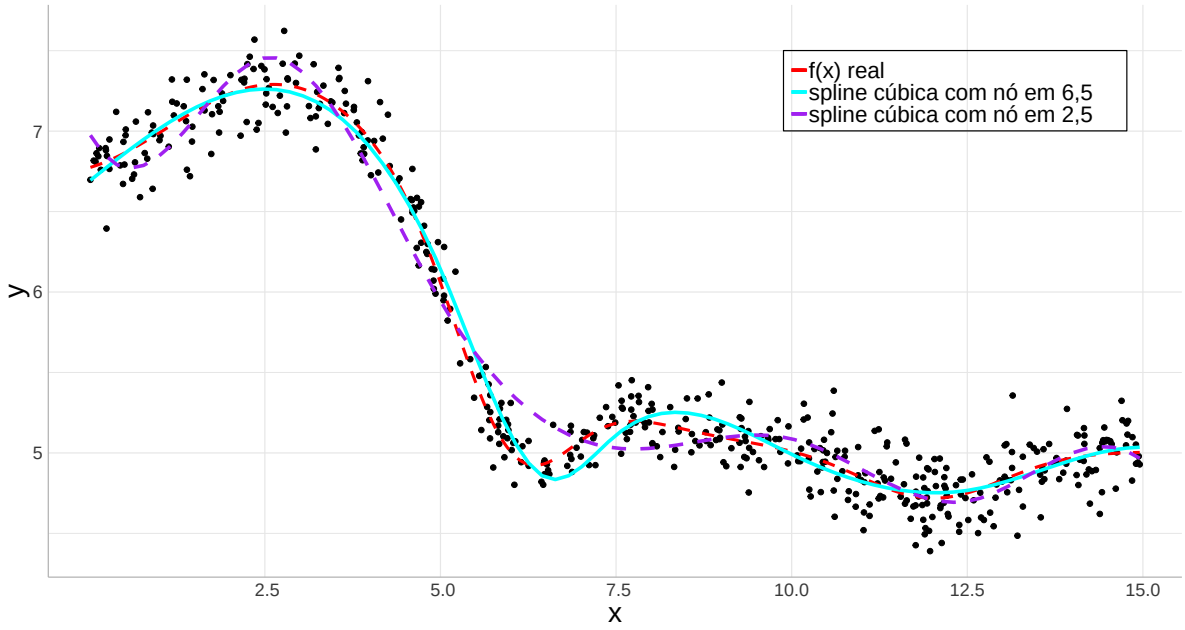


Figura 12: Curva estimada da função suavizada $f(x)$ para os dados.

2.4.2 *Smoothing Splines* (*Splines* Suavizadores)

Quando se pensa em ajustar uma curva suavizada para os dados, pode-se pensar em uma função f que minimiza a soma dos quadrados dos resíduos. Pode-se ter problemas com essa abordagem, pois sem os devidos cuidados, a soma dos quadrados dos resíduos pode ser zerada por tal função e, assim, obter-se uma função f que consegue interpolar todos os pontos y_i. Isso representaria um sobreajuste dos dados, resultando em uma curva extremamente flexível, o que não é o objetivo. Então, o que se busca é uma função que resulte em uma soma dos quadrados dos resíduos pequena, mantendo-se suave. Uma possibilidade para se atingir esse objetivo é obter a função f que minimize

$$\sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \int f''(x)^2 dx, \tag{2.12}$$

onde λ é uma constante não-negativa representando um parâmetro de suavização. O primeiro termo mede o quão próxima a curva está dos dados e o segundo termo adiciona uma penalização à curvatura da função. A segunda derivada $f''(x)$ indica o quanto a inclinação da curva vai mudando e a integração da segunda derivada ao quadrado no domínio da função dá uma medida da curvatura agregada nesta função. Ou seja, a penalidade tem valores altos quando f oscila bastante e quando f é mais suave a penalidade é pequena. O parâmetro de suavização, λ, estabelece o balanço entre o quão suave a curva é e o quão próxima aos dados ela está. Quanto maior o valor de λ, mais suave a curva é. À medida

que λ se aproxima de zero, a penalidade vai perdendo sua importância e a estimação será uma função extremamente flexível, duas vezes diferenciável que interpola os dados, oscilando muito. À medida que λ cresce, tendendo à infinito, a função vai se tornando cada mais suave até se tornar uma reta passando o mais próximo possível dos pontos y_i (GARETH et al., 2013).

O *spline* suavizador, ou *smoothing spline*, é essa função f que minimiza a Equação (2.12). Sabe-se que se trata de um polinômio cúbico "por partes" *(piecewise-cubic polynomial)* com nós definidos nos valores únicos de x_i, $i = 1, \ldots, n$, com primeira e segunda derivadas contínuas nestes nós e que é linear para valores que se encontram fora do intervalo entre x_1 e x_n. Essa quantidade de nós e, consequentemente, de graus de liberdade pode parecer excessiva, porém, o valor do parâmetro de suavização restringe esses graus de liberdade severamente, à medida que λ cresce.

A curva f pode ser escrita na forma da Equação (2.11), então, é linear nos parâmetros, γ_j. Dessa forma, a penalização pode ser escrita na forma quadrática com representação matricial da seguinte forma:

$$f''(x) = \boldsymbol{\gamma}^T \boldsymbol{D} \Longrightarrow [f''(x)]^2 = \boldsymbol{\gamma}^T \boldsymbol{D}\boldsymbol{D}^T \boldsymbol{\gamma},$$

onde $\boldsymbol{D}$ é o vetor com os valores $b_j''(x)$ e $\boldsymbol{\gamma}$ é o vetor dos parâmetros do modelo. Substituindo $\boldsymbol{D}\boldsymbol{D}^T = \boldsymbol{S}$,

$$\int f''(x)^2 dx = \boldsymbol{\gamma}^T \boldsymbol{S} \boldsymbol{\gamma} \tag{2.13}$$

.

Agora, a estimação do modelo por mínimos quadrados penalizado se dá minimizando

$$||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\gamma}||^2 + \lambda \boldsymbol{\gamma}^T \boldsymbol{S} \boldsymbol{\gamma} \tag{2.14}$$

em relação à $\boldsymbol{\gamma}$. Para se estimar o nível de suavização do modelo é necessário estimar o parâmetro de suavização λ.

Caso seja fixado um valor para λ, a expressão para os valores que minimizam a Equação (2.14) é

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{S})^{-1} \boldsymbol{X}^T \boldsymbol{y}. \tag{2.15}$$

Isso implica que ao mudar o valor de λ, muda-se o modelo através de diferentes níveis de suavizações. Na prática, esse valor de λ não é conhecido e precisa ser estimado. Na seção 2.4.3 é apresentada a forma que se define λ. Esses resultados e mais detalhes podem ser encontrados no livro de Wood (2017).

2.4.3 Escolha do parâmetro de suavização

É necessário que haja cuidado para que os dados não sejam suavizados demais ou de menos e isso depende da escolha do valor do parâmetro de suavização, λ. Seu valor precisa manter a curva estimada, $\hat{f}$, o mais próximo possível do seu valor real, f, desconhecido.

Existem algumas abordagens para a escolha do λ e a utilizada neste trabalho será a da validação cruzada. Mais especificamente a validação cruzada generalizada, onde deseja-se escolher o valor do parâmetro de suavização que minimize o escore da validação cruzada generalizada, ν_g,

$$\nu_g = \frac{n \sum_{i=1}^{n} [y_i - \hat{f}(x_i)]^2}{[n - tr(\boldsymbol{A})]^2}, \tag{2.16}$$

onde $tr(\boldsymbol{A})$ é o traço da matriz chapéu do modelo ajustado para todos os dados.

A validação cruzada generalizada varia da validação cruzada ordinária, em que a ideia é retirar um ponto dos dados por vez, ajustar o modelo para os dados restantes e calcular o quadrado da diferença entre o valor ajustado para esse dado retirado e a resposta desse dado. Em seguida, dividir pela quantidade total de dados para obter uma média. Por fim, o objetivo é escolher o valor de λ que resulta no menor valor do escore, ou seja, escolher o λ que fornece o melhor valor ajustado.

Para maiores detalhes sobre os procedimentos de validação cruzada, ver Wood (2017).

2.4.4 Propriedades da estimação do GAM

O modelo aditivo generalizado é composto pela soma de funções suavizadas, escritas como splines, que compõem seu preditor linear e este, como nos MLG, está relacionado à média da i-ésima observação da variável resposta através de uma função de ligação, monótona e diferenciável. Essa variável resposta pode seguir qualquer distribuição da família exponencial.

No caso deste trabalho, a função de ligação utilizada será a função logística, onde o parâmetro natural da distribuição Bernoulli, $b(\theta)$, visto na Equação (2.2), é usado.

Quando no modelo há mais de uma função suavizada, existe um problema. Como essas funções suavizadas f_j são somadas, há a possibilidade de adicionar uma constante a alguma f_j e, ao mesmo tempo, subtrair essa mesma constante de outra f_k sem haver mudança nas respostas do modelo. Precisa-se aplicar uma restrição antes de ajustar-se o modelo. A restrição de que a função suavizada f_j possui soma zero em seus valores obser-

vados x_{ji}, isto é, $\sum_{i=1}^{n} f_j(x_{ji}) = 0$, é a melhor absorvida pela base via reparametrização.

Resolvido o problema de identificabilidade mencionado acima, pode-se representar o modelo na forma linear e o GAM é ajustado através da maximização da verossimilhança penalizada,

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2}\sum_{j=1}^{p} \lambda_j \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{S_j} \boldsymbol{\beta}, \tag{2.17}$$

onde $\boldsymbol{\beta}$ é o vetor dos coeficientes do modelo contendo os parâmetros correspondentes aos componentes estritamente paramétricos e os parâmetros de cada função suavizada, λ_j é um parâmetro de suavização e $\boldsymbol{S}_j$ é a matriz com as penalidades como bloco diagonal e zeros completando-a. Essa maximização da verossimilhança, na prática, será feita de forma iterativa, até atingir convergência, via algoritmo P-IRLS (*penalized iteratively re-weighted least squares*), mínimos quadrados penalizados iterativamente reponderados, algoritmo similar ao utilizado nos MLG.

2.5 Modelo Proposto

O modelo aditivo generalizado que será utilizado neste trabalho é definido nesta seção. Porém, antes de defini-lo, é importante mencionar que, neste estudo, decidiu-se seguir um caminho de olhar para o problema da modelagem com uma ótica de ocorrência ou não do roubo, tratando a resposta como uma variável binária, principalmente pelo interesse em modelar as probabilidades de ocorrência de roubos e obter um mapa da dinâmica dessas probabilidades. Seria possível, por exemplo, optar por seguir a ótica de contar a quantidade de ocorrências de roubos. Isso, porém, poderia gerar uma grande quantidade de zeros na resposta, o que talvez trouxesse maiores dificuldades para modelar.

Assim, o objetivo é modelar a probabilidade de ocorrência de roubos em uma determinada quadrícula, levando-se em conta as características sociais, demográficas, econômicas, educacionais e geográficas da quadrícula. Pode-se definir, então,

$$Y_i = \begin{cases} 1 & \text{se há ocorrência de roubo na quadrícula } i \\ 0 & \text{se não há ocorrência de roubo na quadrícula } i \end{cases}, \tag{2.18}$$

onde $i = 1, 2, \ldots, n$. E, assim,

$$Y_i \sim Bernoulli(\pi_i), \tag{2.19}$$

onde π_i é a probabilidade de ocorrência dos crimes na quadrícula i, $i = 1, 2, \ldots, n$.

Como visto no Exemplo 2.1, a distribuição Bernoulli pertence à família exponencial na forma canônica e a função de ligação utilizada será o parâmetro natural da distribuição, $b(\theta)$, da Equação (2.2):

$$g(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right). \tag{2.20}$$

Dessa forma, a partir da Equação (2.10), o modelo proposto será:

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{l=1}^{L}\beta_l C_{l,i} + \sum_{q=1}^{Q} f_q(M_{q,i}) + f_{Q+1}(longitude_i; latitude_i), \tag{2.21}$$

onde:

- C_l são as covariáveis que possuem relação linear com a resposta, $l = 1, \ldots, L$;
- M_q são as covariáveis que possuem relação não-linear com a resposta, $q = 1, \ldots, Q$;
- $f_q(.)$ são funções suavizadas via *spline* cúbica das covariáveis, $q = 1, \ldots, Q$;
- $f_{Q+1}(longitude_i; latitude_i)$ representa a função suavizada da longitude e latitude do centroide da quadrícula i, $i = 1, \ldots, n$.

3 Análise dos Resultados

Neste capítulo, apresenta-se as análises e resultados da utilização do modelo proposto na Seção 2.5 para os dados de roubos no município do Rio de Janeiro. Para tal, o capítulo divide-se em três seções. Na primeira, é explicada a manipulação e tratamento dos dados obtidos para o estudo e definidos na Seção 2.2. Na segunda, faz-se uma análise descritiva dos dados tratados para maior compreensão dos mesmos. Na última parte, os resultados do modelo escolhido para o estudo são mostrados e discutidos.

Toda manipulação de dados, gráficos, mapas e análises foram feitas no *software* R (R Core Team, 2022).

3.1 Limpeza e Manipulação dos Dados

Em relação às informações dos roubos no município do Rio de Janeiro, para análise neste trabalho, foram selecionados e agrupados os dados referentes aos meses de janeiro a dezembro do ano de 2019, por ser o ano mais recente que se tem informação na base do ISP (Seção 2.2.1), sem a influência da pandemia de Covid-19. Ou seja, não houve variação temporal nas análises feitas.

Como foi explicado na Seção 2.2, houve necessidade de manipulação dos dados para que fossem criadas as variáveis do trabalho. O banco de dados do Censo Demográfico foi obtido pela união de diversas planilhas com informações agrupadas de acordo com o critério próprio do IBGE. As informações que eram necessárias para a criação das variáveis presentes neste trabalho foram identificadas, e uma nova base de dados foi produzida com as variáveis criadas para os indicadores do censo demográfico. O banco de dados do ISP possuía informações sobre diversos crimes de forma codificada. Os códigos para a definição de roubo deste trabalho, apresentado na Seção 1.3, foram identificados e os dados, filtrados. Precisou haver nova filtragem, para que apenas quadrículas pertencentes à cidade do Rio de Janeiro estivessem presentes. O mesmo aconteceu com os dados do censo demográfico (Seção 2.2.2.1). As manipulações descritas na Seção 2.2.1 foram, então,

feitas para obtenção das variáveis longitude e latitude dos centroides das quadrículas.

Foi necessário, também, obter os dados com os polígonos georreferenciados dos limites dos setores censitários e filtrá-los para que apenas os 10.233 setores que se tinha dados estivessem presentes. Dessa forma foi possível fazer o cruzamento e a interseção com os pontos georreferenciados dos centroides das 31.508 quadrículas e determinar o setor censitário ao qual cada quadrícula faz parte. Como os setores censitários com dados disponíveis não contemplam áreas que não são habitadas, como áreas verdes, lagoas e algumas ilhas, e a malha de quadrículas cobre toda a extensão do município, inclusive essas áreas não habitadas, houve remoção de centroides por não pertencerem a setor algum.

Sabendo o setor censitário ao qual cada quadrícula faz parte, foi feita uma agregação de dados de forma que as quadrículas recebessem os dados dos indicadores do censo demográfico de acordo com seus setores censitários. Foi constatado, em um momento seguinte, que ainda havia 75 setores censitários com dados faltantes e os centroides das quadrículas que pertenciam a esses setores também foram removidos.

Através de análises em mapas gerados com os polígonos das quadrículas e dos setores censitários, percebeu-se que havia quadrículas que cobriam outras áreas não habitadas, como áreas agrícolas, de indústrias, militares ou apenas campos, como por exemplo a base aérea de Santa Cruz e a Restinga de Marambaia. Os centroides dessas quadrículas também foram removidos por não serem de interesse do estudo, dado que não há circulação relevante de pessoas por essas áreas, e poderiam causar um efeito de confusão no modelo. Dessa forma, o número de quadrículas, ou seja, de centroides na base de dados final foi de 20.667. Essas quadrículas estão representadas no mapa da Figura 13.

Definidos os centroides das quadrículas para fazer parte da base de dados final, através de seus pontos georreferenciados, as distâncias descritas nas Seções 2.2.2.2 e 2.2.2.3 foram, então, calculadas e adicionadas à base. Assim, foi formada a base de dados final do trabalho.

3.2 Análise Descritiva

Nesta seção são feitas algumas análises descritivas, com foco espacial, das variáveis selecionadas para o estudo de forma a entender melhor como cada uma se comporta e tentar identificar padrões entre elas.

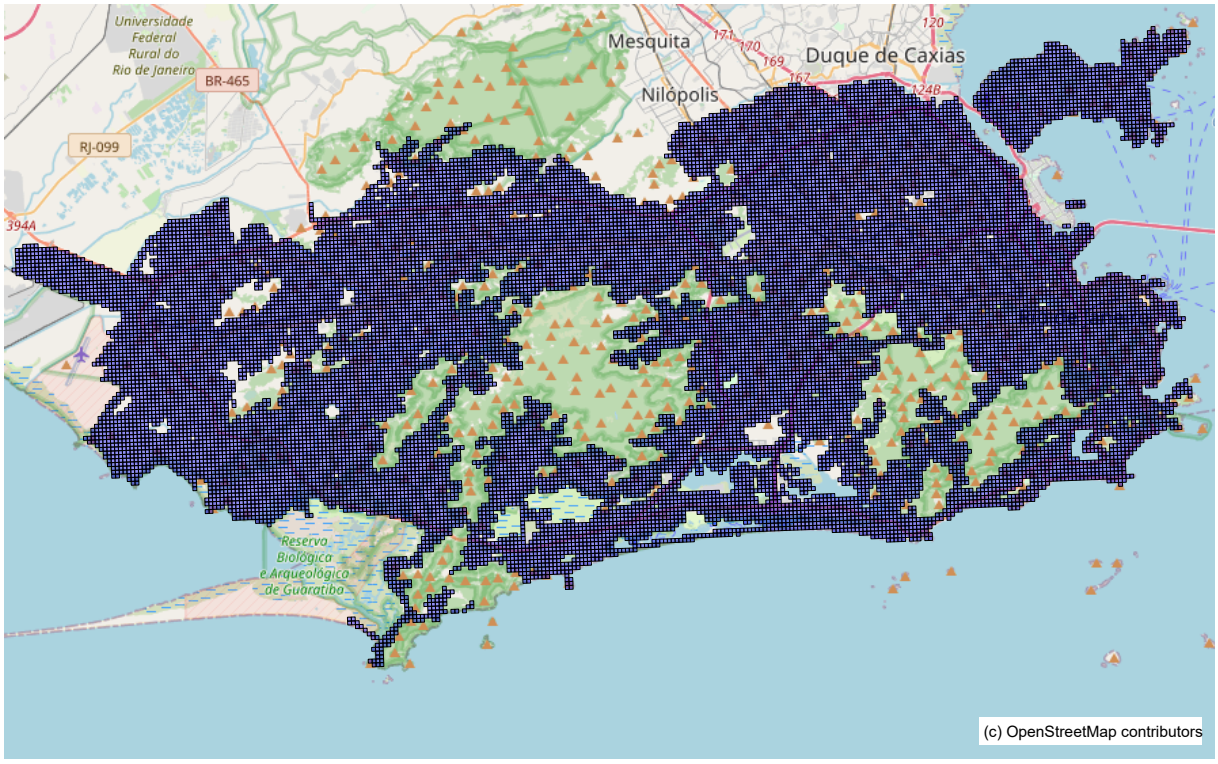


Figura 13: 20.667 quadrículas após limpeza e manipulação no mapa do município do Rio de Janeiro.

A variável de interesse é a ocorrência de roubo na quadrícula i, $i = 1, \ldots, 20.667$. Ocorreram roubos em 8.841 quadrículas enquanto em 11.826, não houve roubos. Em percentuais, esses valores correspondem à $42,78\%$ e $57,22\%$, respectivamente. A disposição espacial da ocorrência de roubos nas quadrículas está exposta na Figura 14.

Pode-se perceber que há maior concentração de roubos nas zonas norte, central e sul da cidade que são regiões mais urbanas, possuem um trânsito maior de pessoas no dia-a-dia e possuem densidade demográfica maior que a zona oeste. A quadrícula que teve o maior número de roubos se localiza em Barros Filho, próxima a interseção entre Estrada João Paulo e Avenida Brasil, com 155 roubos registrados no ano de 2019. O total de roubos no município no ano de 2019 foi de 70.791 e média do número de roubos por quadrícula foi de $3,43$. Ao todo, 4.914 quadrículas estão acima da média de roubos e $76,22\%$ das quadrículas têm entre 0 e 3 roubos registrados em 2019.

As variáveis explicativas em que são medidas distâncias entre o centroide da quadrícula e um local geográfico determinado não dependem do setor censitário, assim, podem variar de acordo com cada centroide, diferentemente das variáveis explicativas que provém do censo demográfico, que variam em grupo de acordo com a vizinhança (setor censitário) de um determinado centroide. As variáveis explicativas de distância e algumas medidas

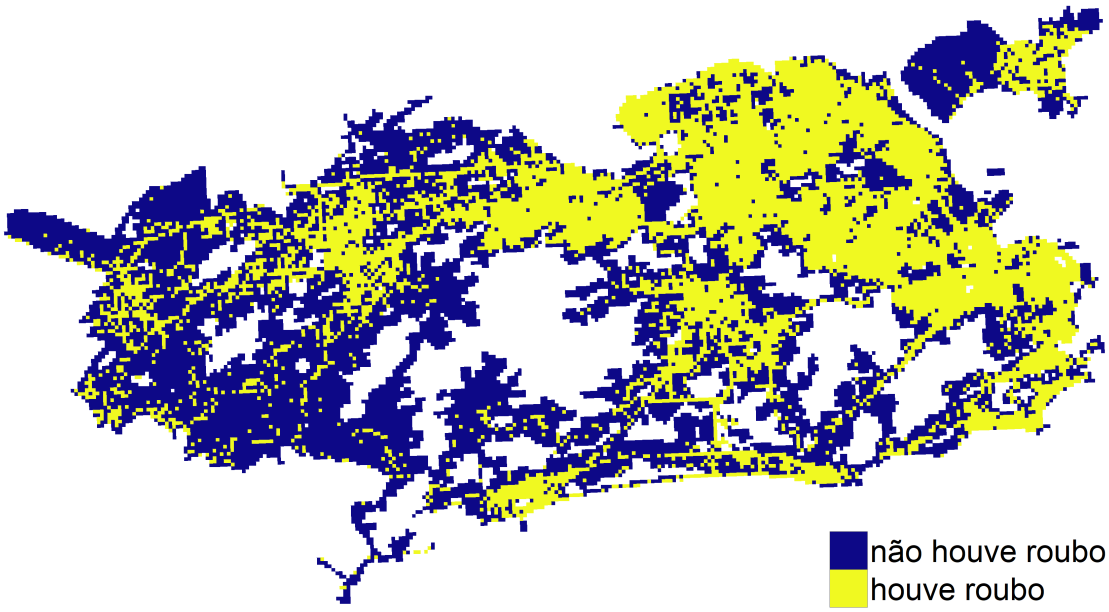


Figura 14: Mapa da variável de interesse, roubo, segundo as quadrículas.

resumo são mostradas na Tabela 1. Observa-se que, dentre os centroides de áreas habitadas que foram estudados, nenhum possui distância maior do que $4,2$ quilômetros até uma favela, mostrando que essas áreas mais pobres e marginalizadas pelo poder público possuem presença considerável no território da cidade. A mesma situação ocorre em relação às escolas e unidades de saúde, indicando uma maior presença do estado. Nota-se que todos os centroides estão em um raio de $3,41$ quilômetros de alguma escola ou de menos de 5 quilômetros de uma unidade de saúde.

Tabela 1: Medidas resumo para as variáveis que medem distâncias (em km)

Variável	Mínimo	1º Quartil	Mediana	3º Quartil	Máximo	Média
dist_fav	0,00	0,15	0,38	0,74	4,18	0,55
dist_esc	0,00	0,24	0,42	0,80	3,41	0,60
dist_us	0,00	0,49	0,83	1,40	4,96	1,05
dist_trem	0,00	1,29	3,49	7,47	18,58	4,78
dist_via	0,00	0,12	0,34	0,78	5,92	0,64
dist_brt	0.01	0.97	2.21	4.14	9.86	2.75

Para deixar mais claras essas observações, a Figura 15 mostra, agora, os mapas com as distâncias padronizadas entre os centroides e os locais geográficos estudados. A padronização se dá dividindo cada valor i das variáveis explicativas pelo valor máximo encontrado em cada uma delas. Dessa forma, tem-se uma escala homogênea, de 0 a 1,

assim como nas proporções dos indicadores do censo demográfico. Ou seja, dessa maneira, compara-se todas as variáveis na mesma escala. As áreas com tons mais frios e escuros estão mais próximas das localidades geográficas em estudo e as áreas com tons mais quentes e claros estão mais distantes.

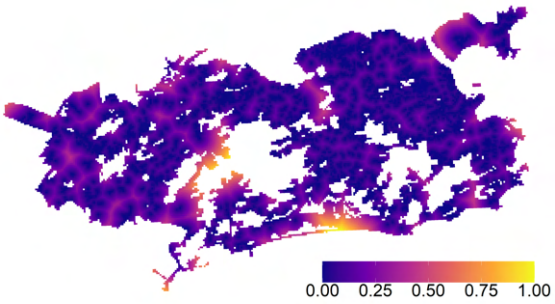
A Figura 15(e) é a que aparenta ter um padrão mais próximo do padrão de roubos da Figura 14, podendo indicar que os roubos ocorrem mais frequentemente mais próximos dessas vias. Nas Figuras 15(a), 15(b) e 15(c) pode-se ver que apesar de haver muitos dos locais espalhados pela cidade, e das distâncias não serem grandes, existe uma concentração menor na zona oeste do que nas outras e como consequência as distâncias nessa região tendem a ser maiores. De uma forma geral pode-se ver um padrão similar ao da ocorrência de roubos. As figuras que menos se assemelham a esse padrão são as Figuras 15(d) e 15(f). As linhas do trem viajam da região do centro da cidade em direção ao norte e oeste, assim, áreas mais ao sul do território da cidade, como Barra da Tijuca e Recreio dos Bandeirantes, possuem longas distâncias até as estações. As linhas do BRT estão mais presentes nas zonas norte e oeste, deixando a zona sul com as maiores distâncias até as estações. Há áreas ao norte do mapa, na zona oeste, que possuem as maiores distâncias, também.

Na Tabela 2 encontram-se medidas resumo das variáveis explicativas construídas a partir do censo demográfico referentes às proporções de domicílios com certas características, conexão à rede de água, de esgoto, presença de banheiro exclusivo dos moradores, energia elétrica, coleta de lixo e ausência de moradores do sexo masculino, conforme descrito na Seção 2.2.2.1. Os mapas na Figura 16 ilustram essas informações de forma espacial.

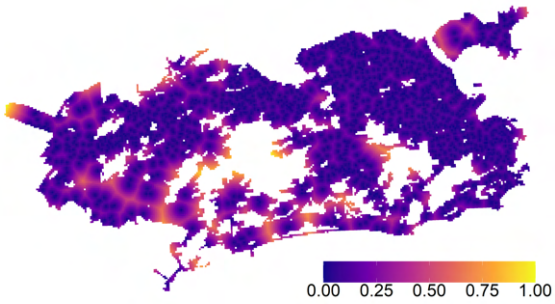
Tabela 2: Medidas resumo para as variáveis de proporção de domicílios

Variável	Mínimo	1º Quartil	Mediana	3º Quartil	Máximo	Média
$D101$ (água)	0,0000	0,9510	1,0000	1,0000	1,0000	0,9103
$D102$ (banheiro)	0,8228	1,0000	1,0000	1,0000	1,0000	0,9984
$D103$ (esgoto)	0,0000	0,4331	0,9283	0,9958	1,0000	0,7177
$D104$ (coleta)	0,0000	0,9924	1,0000	1,0000	1,0000	0,9813
$D105$ (energia)	0,8590	1,0000	1,0000	1,0000	1,0000	0,9991
$D106$ (sem masc.)	0,0000	0,0924	0,1241	0,1622	0,5000	0,1312

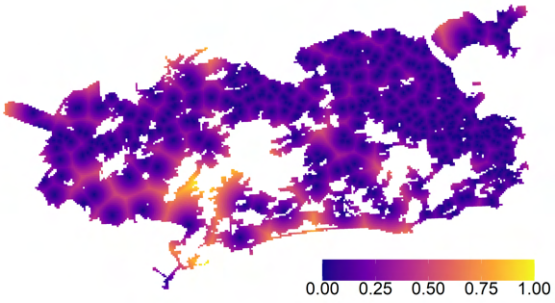
As variáveis *proporção de domicílios com energia elétrica*, *proporção de domicílios com lixo coletado* e *proporção de domicílios com banheiro exclusivo* possuem o valor 1 para quase todas as quadrículas, fato mostrado, também, nas Figuras 16(e), 16(d) e 16(b), onde os mapas são quase 100% de cor amarela. A variável *proporção de domicílios com abastecimento de água da rede geral*, como mostra a Figura 16(a), também possui um



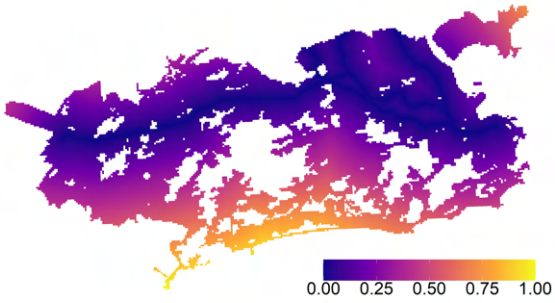
(a) distância até a favela mais próxima



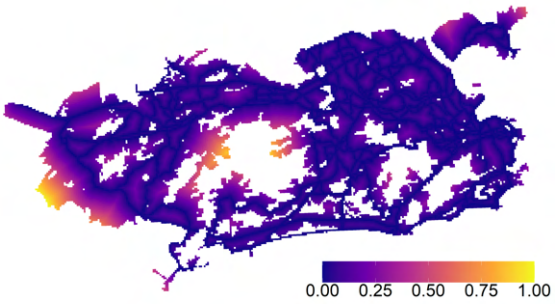
(b) distância até a escola mais próxima



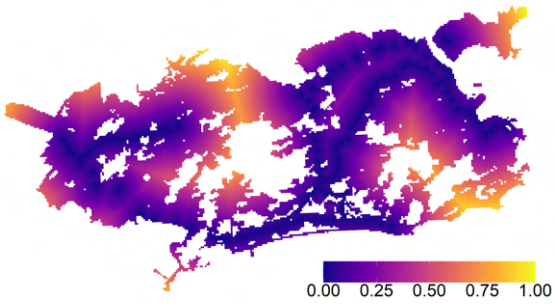
(c) distância até a unidade de saúde mais próxima



(d) distância até a estação de trem mais próxima



(e) distância até a via mais próxima



(f) distância até a estação do BRT mais próxima

Figura 15: Mapa das variáveis de distâncias padronizadas

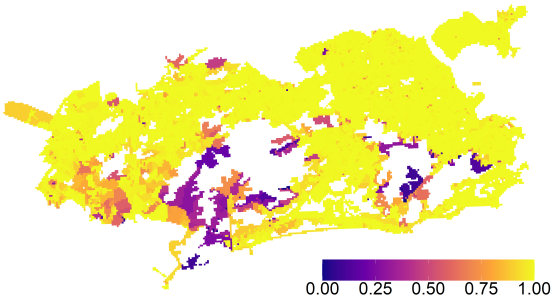
mapa predominantemente amarelo, com algumas células da zona oeste, na região de Vargem Grande, Vargem Pequena, Grumari e Itanhangá, áreas menos populosas, possuindo proporções pequenas.

A *proporção de domicílios com esgotamento da rede geral*, também, possui um mapa, predominantemente amarelo e com proporções menores na zona oeste, em sua maioria, e alguns pontos da zona norte, Figura 16(c). Porém, a Tabela 2 mostra que ainda há 25% das células com proporções abaixo de $0,4331$. Já a *proporção de domicílios sem morador do sexo masculino*, Figura 16(f) e Tabela 2, apesar de possuir uma média baixa e 75% das células com proporções abaixo de $16,22\%$, as áreas onde as proporções sobem um pouco, com cores próximas ao lilás, se assemelham bastante às áreas com incidência de roubos mostradas na Figura 14.

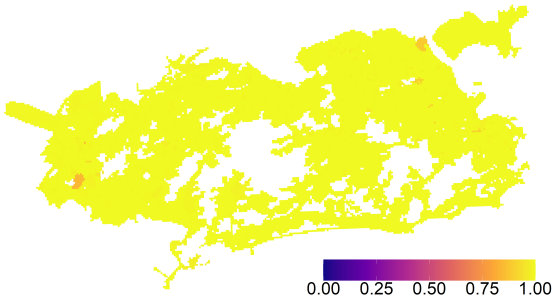
As proporções de domicílios segundo características de seus entornos estão ilustradas na Figura 17 e algumas de suas medidas resumo encontram-se na Tabela 3. Como no caso da Figura 16(a), as Figuras 17(a), 17(e) e 17(f) também apresentam proporções predominantemente em um extremo. As duas primeiras no 0 e a última no 1. Existem células que apresentam proporções fora dos extremos, e até no extremo oposto, por todas as regiões da cidade, porém, um grupo pequeno. Não é esperado que essas variáveis tragam influência positiva ao modelo.

Os padrões espaciais apresentados nas Figuras 17(b) e 17(c) são bastante próximos do padrão de ocorrência de roubos da Figura 14. As áreas próximas a Vargem Grande, Guaratiba, Sepetiba e Santa Cruz, que são regiões menos urbanizadas, apresentam uma grande concentração de células com proporções baixas de calçamento e pavimentação. Outras concentrações menores podem ser observadas na região do Caju, Manguinhos, Maré e Complexo do Alemão, que constituem regiões mais pobres da cidade. As áreas referentes à *proporção de domicílios onde existe arborização*, na Figura 17(d), apresentam um padrão ligeiramente semelhante, mas mostram uma maior variabilidade de proporções ao longo do mapa da cidade, o que pode ser notado através da maior diversidade de cores que se vê nessa figura.

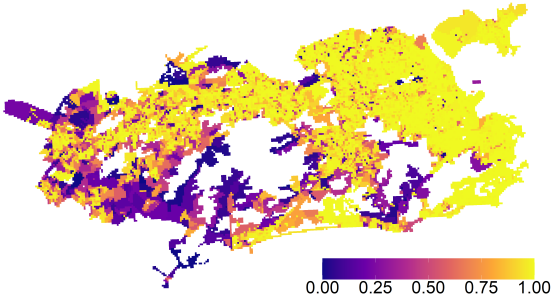
Os indicadores construídos com as características da população e renda foram: *proporção de pessoas residentes de cor ou raça branca*, *proporção de pessoas com menos de 18 anos de idade*, *proporção de pessoas com mais de 59 anos de idade*, *população média por quadrícula*, *proporção de domicílios com rendimento nominal mensal domiciliar per capita de até 1 salário mínimo* e *proporção de domicílios com rendimento nominal mensal domiciliar per capita de mais de 10 salários mínimos*. Assim como as variáveis de



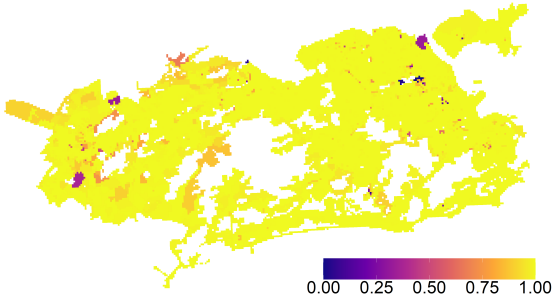
(a) prop. domicílios com abastecimento de água da rede



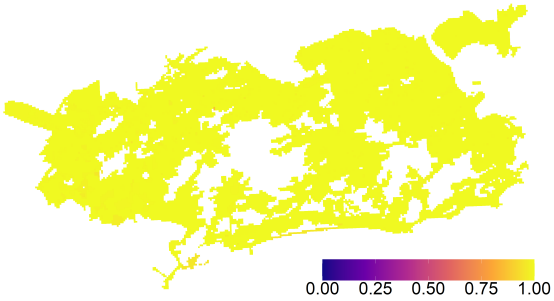
(b) prop. domicílios com banheiro de uso exclusivo



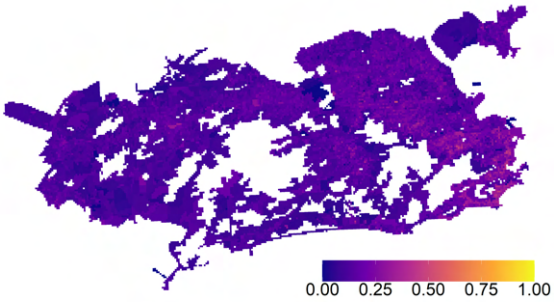
(c) prop. domicílios com esgotamento da rede geral



(d) prop. domicílios com lixo coletado

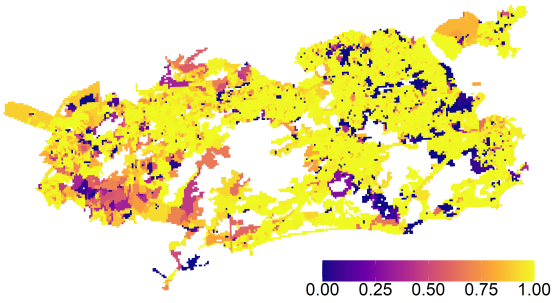


(e) prop. domicílios com energia elétrica

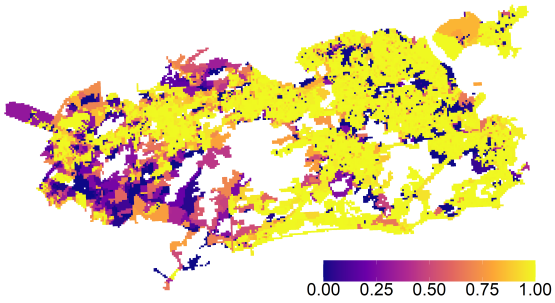


(f) prop. domicílios sem morador do sexo masculino

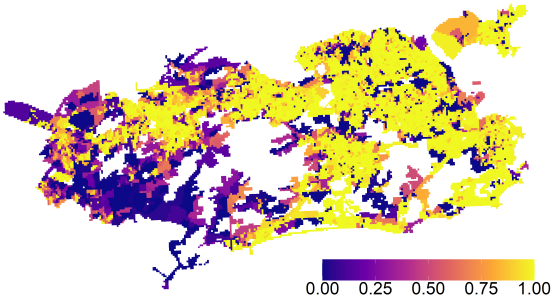
Figura 16: Mapa das variáveis de proporções de domicílios e suas características



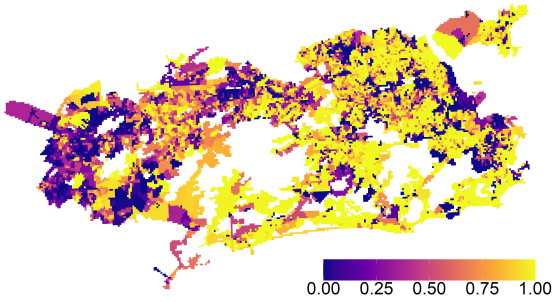
(a) prop. domicílios onde existe iluminação pública



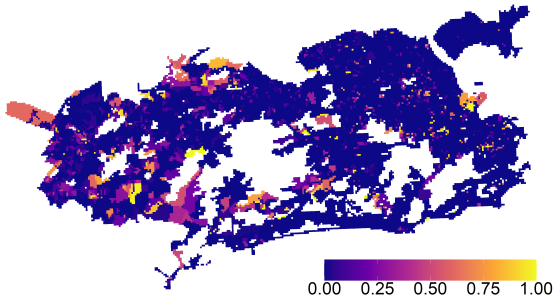
(b) prop. domicílios onde existe pavimentação



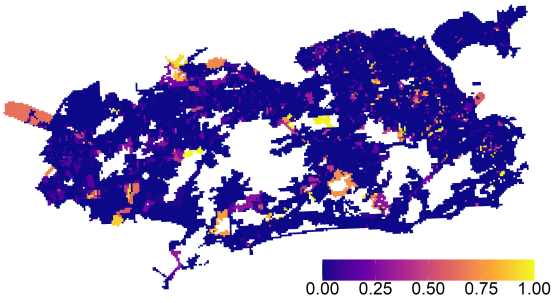
(c) prop. domicílios onde existe calçada



(d) prop. domicílios onde existe arborização



(e) prop. domicílios onde existe esgoto a céu aberto



(f) prop. domicílios onde existe lixo acumulado

Figura 17: Mapa das variáveis de proporções de domicílios segundo características de seus entornos

Tabela 3: Medidas resumo para as variáveis de proporção de domicílios segundo seus entornos

Variável	Mínimo	1º Quartil	Mediana	3º Quartil	Máximo	Média
$E101$ (iluminação)	0,0000	0,8240	0,9831	1,0000	1,0000	0,8302
$E102$ (pavimentação)	0,0000	0,5241	0,9643	1,0000	1,0000	0,7390
$E103$ (calçada)	0,0000	0,1393	0,8358	0,9959	1,0000	0,6128
$E104$ (arborização)	0,0000	0,3192	0,8043	0,9927	1.0000	0,6457
$E105$ (esgoto aberto)	0,0000	0,0000	0,0000	0,0176	1,0000	0,0824
$E106$ (lixo acum.)	0,0000	0,0000	0,0000	0,0000	1,0000	0,0632

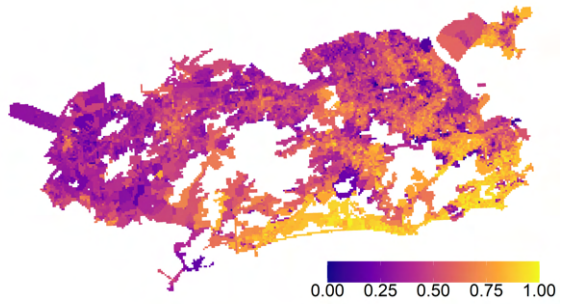
Tabela 4: Medidas resumo para as variáveis de características da população e renda

Variável	Mínimo	1º Quartil	Mediana	3º Quartil	Máximo	Média
$P301$ (branca)	0.0206	0.3839	0.4946	0.6501	0.9865	0.5226
$P1301$ (menos 18)	0.0000	0.2024	0.2491	0.2931	0.5479	0.2521
$P1302$ (mais 59)	0.0000	0.0864	0.1252	0.1725	0.6471	0.1346
$PQ1$ (pop. média)	0.82	21.58	109.50	334.50	3388.00	214.92
$DR01$ (renda 1-)	0.0000	0.2553	0.4685	0.6622	1.0000	0.4600
$DR02$ (renda 10+)	0.0000	0.0000	0.0044	0.0255	0.7784	0.0446

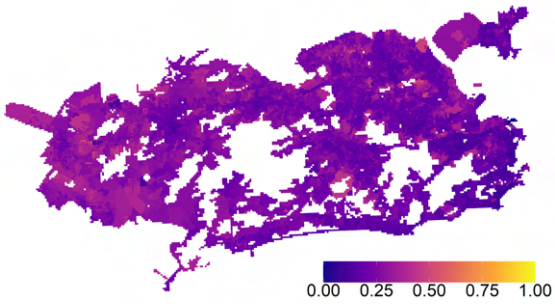
distância, para a exibição no mapa, a *população média por quadrícula* foi padronizada dividindo-se cada valor i pelo valor máximo encontrado para ela, de forma que se compare todas as variáveis na mesma escala. Pode-se analisar as características espaciais da variáveis mencionadas através dos mapas da Figura 18 e detalha-se algumas medidas resumo na Tabela 4.

As proporções de população branca crescem ao mover-se em direção ao sul e sudeste do território municipal, como mostra a Figura 18(a). A zona sul, Barra da Tijuca e Recreio dos Bandeirantes são as áreas onde exitem as maiores proporções da cidade. Já as menores proporções se dão ao norte da zona oeste e na zona norte. As variáveis de renda se associam com as proporções de população branca, de forma que as maiores rendas, mostradas na Figura 18(f), se encontram nas mesmas áreas de maiores proporções de população branca. E segue que as áreas mais pobres, mostradas na Figura 18(e), coincidem com as áreas de menores proporções de população branca.

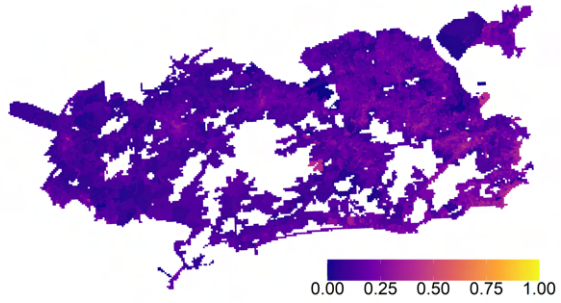
O padrão espacial das proporções das variáveis de idade, maiores do que 59 anos, Figura 18(c), e menores do que 18, 18(c), se mostram opostos. Regiões com maiores proporções de população com a faixa de idade mais baixa possuem menores proporções de população com a faixa de idade mais alta. A região central da cidade e a zona sul são onde as proporções de população com mais de 59 anos de idade são maiores, também, onde as proporções de população com menos de 18 anos são menores. Nas zonas oeste e norte são onde encontra-se as maiores proporções de população com menos de 18 anos e



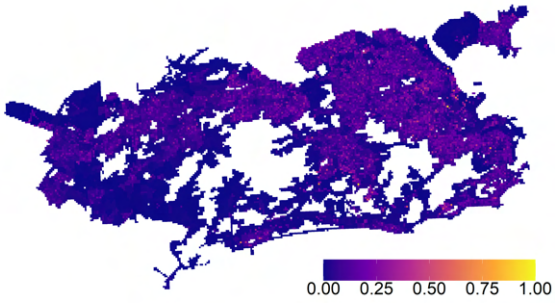
(a) prop. de residentes de cor ou raça branca



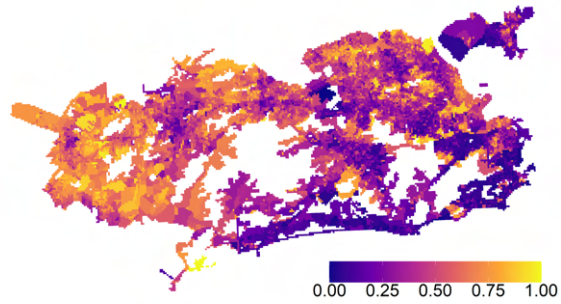
(b) prop. de pessoas com menos de 18 anos



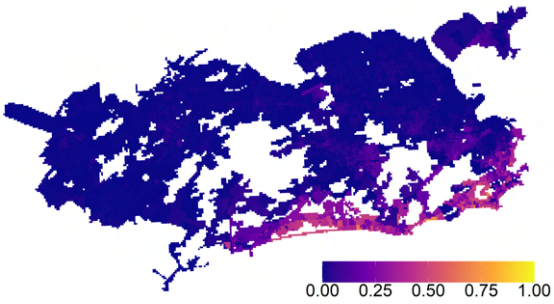
(c) prop. de pessoas com mais de 59 anos



(d) população média por quadrícula



(e) prop. dom. renda até 1 salário



(f) prop. dom. renda maior que 10 salários

Figura 18: Mapa das variáveis de características da população e renda

mais de 59 anos.

O padrão espacial da variável *população média por quadrícula*, mostrado na Figura 18(d), é semelhante ao padrão de ocorrência de roubos da Figura 14. Tal fato pode indicar uma importante contribuição dessa variável para o modelo. Seu valor máximo ocorre em Gardênia Azul e seu valor mínimo ocorre em Santa Teresa, na área próxima ao Cristo Redentor.

3.3 Resultados do Modelo

Aqui são apresentados os resultados do ajuste do modelo proposto na Equação (2.21). Para modelar a probabilidade da ocorrência de roubo em uma determinada quadrícula, utilizando modelos aditivos generalizados, definiu-se, na Equação (2.18), a variável de interesse como a ocorrência ou não de roubo nessa quadrícula no ano de 2019. Foram associadas à essa variável, algumas características sociais, demográficas, econômicas e geográficas das quadrículas, descritas na Seção 2.2, para tentar compreender o comportamento dos roubos no município do Rio de Janeiro.

Não foram utilizados métodos de seleção de variáveis para este estudo. As covariáveis foram testadas por meio de inclusão e retirada do modelo. Suas significâncias eram avaliadas através dos gráficos dos seus efeitos parciais e suas influências avaliadas através do acréscimo ou decréscimo da explicação da deviance, que se refere à quanto da variação dos dados o modelo consegue explicar. Os resultados foram sendo comparados para que se chegasse no melhor modelo, com um maior poder de explicação, levando-se em conta a interpretação e simplicidade do mesmo.

As covariáveis selecionadas para fazerem parte do modelo final foram: *menor distância entre o centroide da quadrícula i e uma via rápida* ($dist\_via$), *distância do centroide da quadrícula i até a favela mais próxima* ($dist\_fav$), *menor distância entre o centroide da quadrícula i e uma escola* ($dist\_esc$), *menor distância entre o centroide da quadrícula i e uma estação de trem* ($dist\_trem$), *proporção de domicílios particulares permanentes sem morador do sexo masculino* ($D106$), *proporção de domicílios particulares permanentes onde existe calçada* ($E103$), *proporção de pessoas com mais de 59 anos de idade* ($P1302$), *proporção de domicílios particulares com rendimento nominal mensal domiciliar per capita de até 1 salário mínimo* ($DR01$), *população média por quadrícula* ($PQ1$), *longitude do centroide da quadrícula i* (lon) e *latitude do centroide da quadrícula i* (lat). É importante ressaltar que o número de quadrículas sofreu um ajuste de acordo com o que foi descrito

na Seção 3.1.

Assim, a partir da Equação (2.21), o modelo final foi:

$$\begin{aligned} log\left(\frac{\pi_i}{1-\pi_i}\right) =& f_1(dist\_via_i) + f_2(dist\_fav_i) + f_3(dist\_esc_i) \\ &+ f_4(dist\_trem_i) + f_5(D106_i) + f_6(E103_i) + f_7(P1302_i) \\ &+ f_8(DR01_i) + f_9(PQ1_i) + f_{10}(lon_i; lat_i), \end{aligned} \tag{3.1}$$

onde:

- $i = 1, \ldots, 20.667$;
- $f_q(.)$ são funções suavizadas via *spline* cúbica das covariáveis, $q = 1, \ldots, 9$;
- $f_{10}(lon_i; lat_i)$ representa a função suavizada da longitude e latitude do centroide da quadrícula i.

Na Equação 3.1 estão representadas apenas funções suavizadas para todas as covariáveis e isso se dá porque todas elas mostraram possuir relação não-linear com a variável resposta. Ou seja, apesar de tentativas e testes de entrada no modelo de covariáveis com relações lineares com a resposta, nenhum desses testes foi capaz de melhorar o desempenho do modelo. Ademais, todas as covariáveis do modelo final foram significativas com $p\text{-}valor < 0,0001$. O modelo ajustado consegue explicar 46% da variação dos dados, valor obtido da explicação da deviance.

Como dito no parágrafo anterior, obtivemos evidências de que todas as relações entre as covariáveis escolhidas para participarem do modelo final e a resposta não são lineares, o que já justificaria a utilização dos modelos aditivos generalizados em detrimento aos modelos lineares generalizados, por exemplo. Entretanto, para testar a diferença, foi ajustado um MLG utilizando as mesmas covariáveis selecionadas para o modelo final e este obteve um resultado inferior ao GAM, sua explicação da deviance foi de 32%, fato que ratifica a escolha do GAM ao MLG.

Os gráficos das Figuras 19 até 27 mostram os efeitos parciais de cada covariável na variável resposta. Os eixos das ordenadas apontam o valor da função suavizada e estão na escala $log\left(\frac{\pi_i}{1-\pi_i}\right)$. Para converter o valor para a escala de probabilidade π_i, basta fazer a função inversa $\pi_i = \frac{e^{s(.)_i}}{1+e^{s(.)_i}}$. Mostra-se os gráficos desta maneira para facilitar a interpretação da significância das covariáveis, sendo possível verificar mais facilmente quando suas contribuições para o ajuste do modelo são nulas, positivas ou negativas. Os

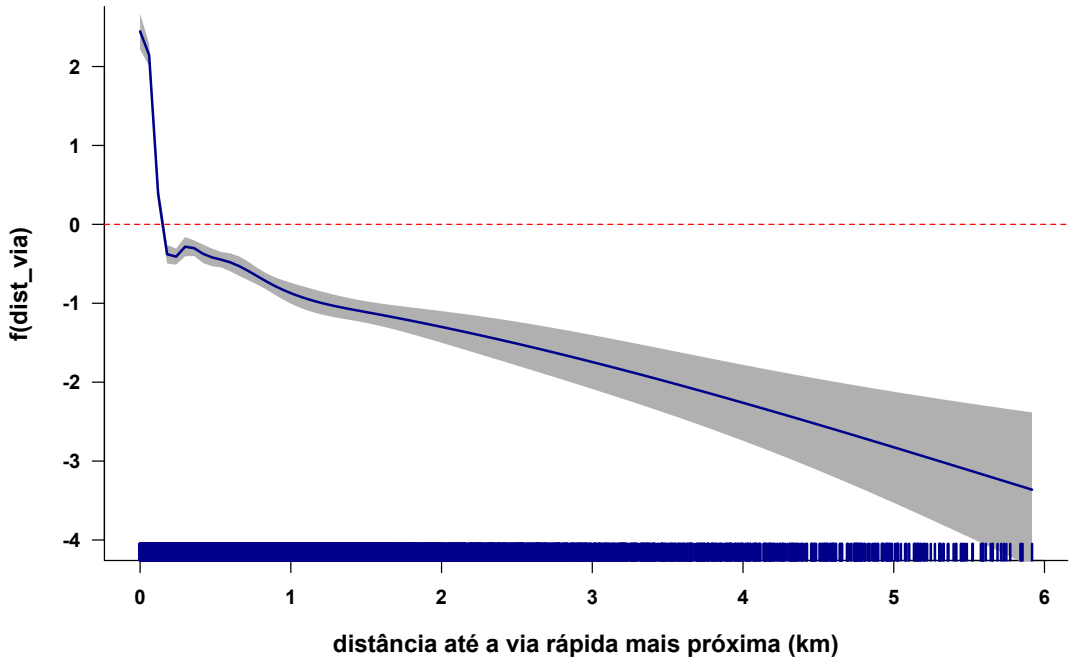


Figura 19: Gráfico do efeito parcial da variável *dist_via*.

sombreados em cinza indicam os intervalos de confiança de 95% para os valores do efeito e o tracejado azul junto ao eixo das abscissas são os valores observados dos dados dos centroides.

Através da Figura 19 vê-se que o que se imaginava sobre os roubos serem mais prováveis de acontecer próximo às vias é verdadeiro. As vias rápidas acabam se tornando rotas de fugas mais fáceis para os criminosos e, assim, à medida que a distância até a via aumenta, o valor do efeito parcial cai. Em um primeiro momento mais rapidamente, entre 0 e $0,2$ quilômetros, e um pouco mais lentamente, a partir de $0,2$. Outro fato importante que se percebe é o alto valor inicial que a função mostra. Quando se está junto à uma via, a contribuição é de $2,56$. Em probabilidade isso é $92,82\%$.

A probabilidade dos roubos acontecerem nos limites das favelas, juntos à elas, é menor do que acontecer mais afastado delas, como mostra a Figura 20. Esse limite de distância, porém, é pequeno. O efeito parcial da variável $dist\_fav$ é negativo entre 0 e $0,15$ quilômetros da favela. Entre $0,15$ e $1,03$, o efeito é positivo indicando maiores probabilidades de ocorrência de roubos. A partir dessa distância, o efeito cai, até perder sua significância, em $3,6$ km.

Distâncias próximas à escolas possuem maiores probabilidades de ocorrência de roubo. Isso se dá, provavelmente, devido à maior concentração e movimentação de pessoas em seus arredores. O efeito positivo inicial se torna negativo em $0,47$ km e, em $2,42$ km

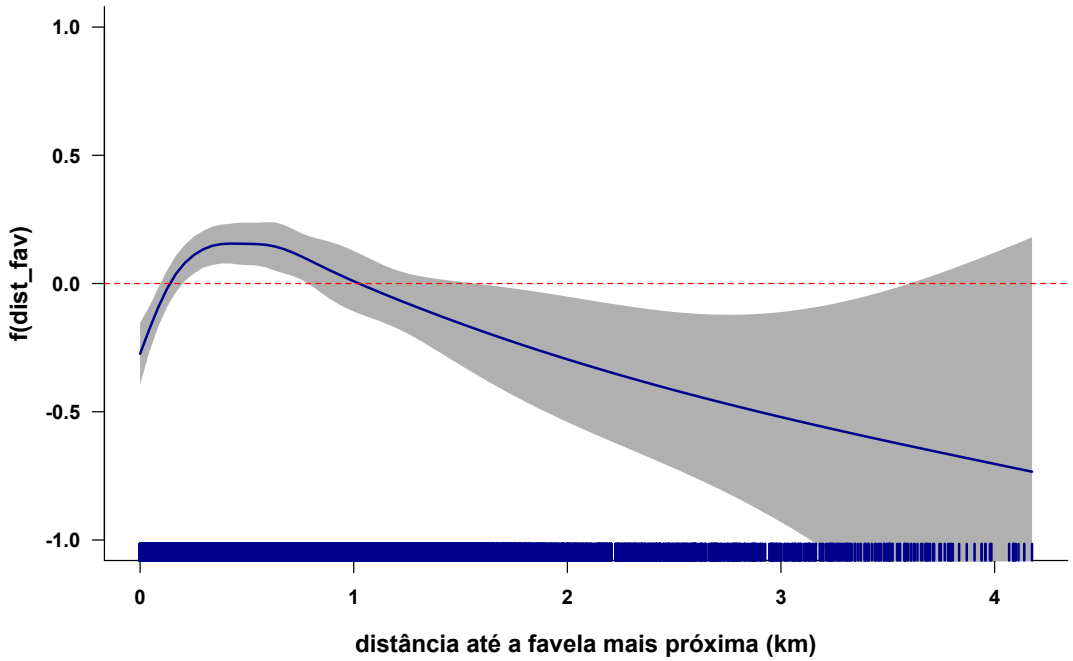


Figura 20: Gráfico do efeito parcial da variável $dist\_fav$.

de distância de uma escola, o efeito perde sua significância, como pode-se constatar na Figura 21.

As estações de trem são um outro exemplo de lugar com alta concentração de pessoas ao redor. Muitas pessoas utilizam esse meio de transporte que faz uma importante ligação entre as zonas oeste e centro/norte. Vê-se na Figura 22 que a variável $dist\_trem$ possui um efeito positivo inicial que vai caindo e torna-se negativo em $3,1$ km e perde significância em $8,9$ km.

É muito comum imaginar que mulheres são mais visadas do que homens para serem vítimas de roubo, assim, foi investigado se regiões com uma quantidade maior de domicílios compostos apenas por mulheres teriam maior probabilidade de roubo. Isso se torna um fato e a confirmação pode ser observada na Figura 23, onde se constata a tendência de crescimento do efeito da variável $D106$ à medida que a proporção de domicílios sem indivíduos masculinos cresce. Percebe-se que as regiões com proporções pequenas, até $16,2\%$, sofrem o efeito negativo e, a partir dessa proporção, o efeito é positivo. Tem-se poucas quadrículas com proporções maiores do que 40% e por isso a largura do intervalo de confiança é muito grande e a variável acaba por perder significância em 45%.

A proporção de domicílios onde tem calçada pode não indicar a influência da presença de calçada em si, mas pode ser um indicativo de urbanização e, assim, faz-se justificada a utilização dessa variável no modelo. Outras variáveis que poderiam indicar urbanização

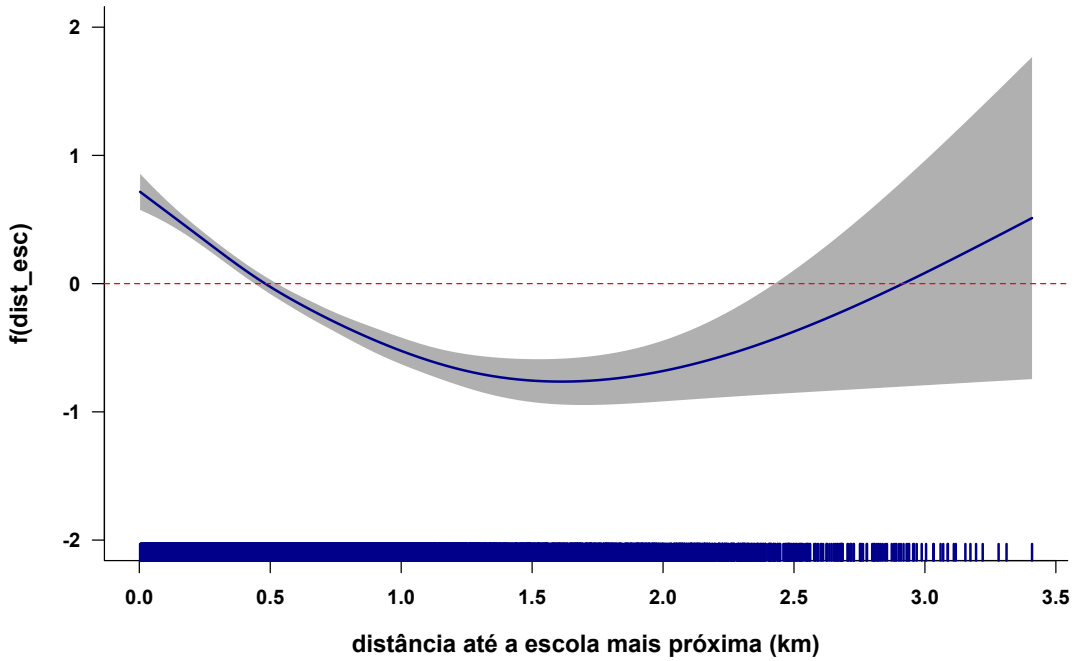


Figura 21: Gráfico do efeito parcial da variável *dist_esc*.

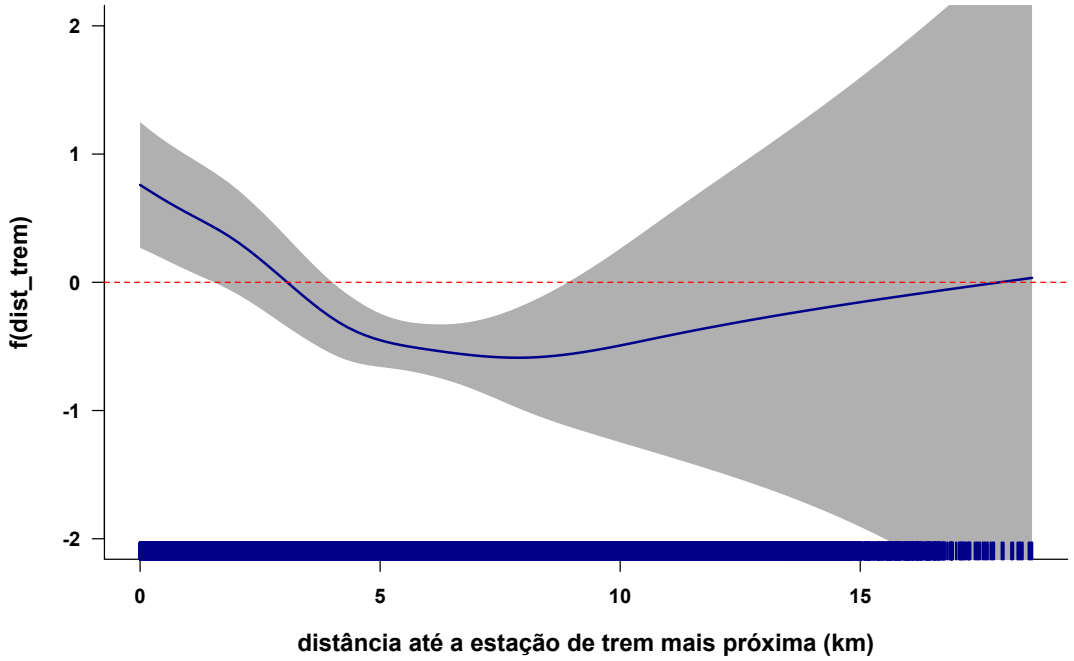


Figura 22: Gráfico do efeito parcial da variável *dist_trem*.

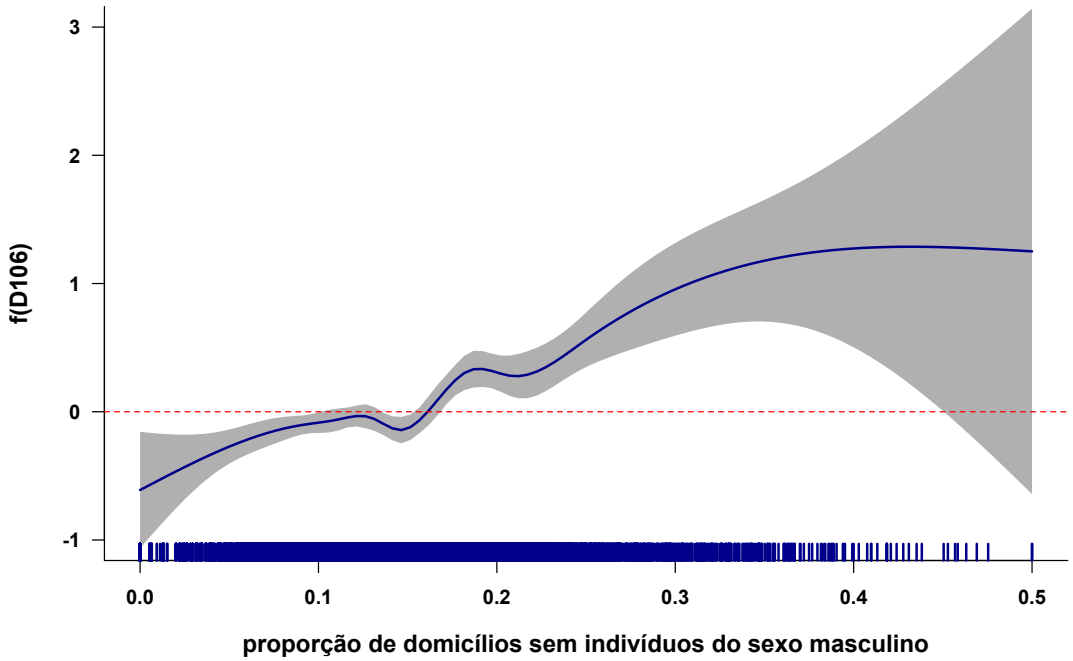


Figura 23: Gráfico do efeito parcial da variável $D106$.

não obtiveram resultados relevantes nos testes dos modelos. Elas não se mostraram significativas ou não faziam sentido para interpretá-las, individualmente, o que poderia confundir os efeitos do modelo. Como a Figura 24 mostra, a interpretação da variável $E103$ é interessante e faz sentido. Regiões com proporções baixas de domicílios com calçamento em seu entorno possuem efeito negativo na probabilidade de ocorrência de roubos e proporções altas possuem efeito positivo. O efeito é negativo até $0,15$, onde a variável passa a ser não significativa, voltando a ter significância quando a proporção está entre $0,6$ e $0,7$ com efeito ainda negativo e passa a ter, novamente, significância em $0,92$, quando começa uma crescente para o efeito positivo dessa variável.

A Figura 25 mostra o efeito parcial da função suavizada da proporção de pessoas idosas por quadrícula na resposta do modelo e constata-se que há maiores probabilidades de ocorrer roubos onde há maiores proporções de idosos. Têm-se efeito negativo nas proporções mais baixas, ate $13,4\%$ e, a partir daí, o efeito positivo cresce até atingir o pico em 29%. Como visto na Tabela 4, temos 25% das proporções entre $17,25\%$, 3^{o} quartil, e $64,71\%$, valor máximo, e temos apenas 368 observações $(1,78\%)$ com proporções maiores que 30%. Dessa forma, talvez, os valores dos efeitos negativos nas proporções acima desse valor se devam à influência de outros fatores, alheios à variável $P1302$.

Analisando a Figura 26, onde mostra-se o efeito parcial da variável proporção de domicílios com renda nominal mensal per capita de até 1 salário mínimo no ajuste do

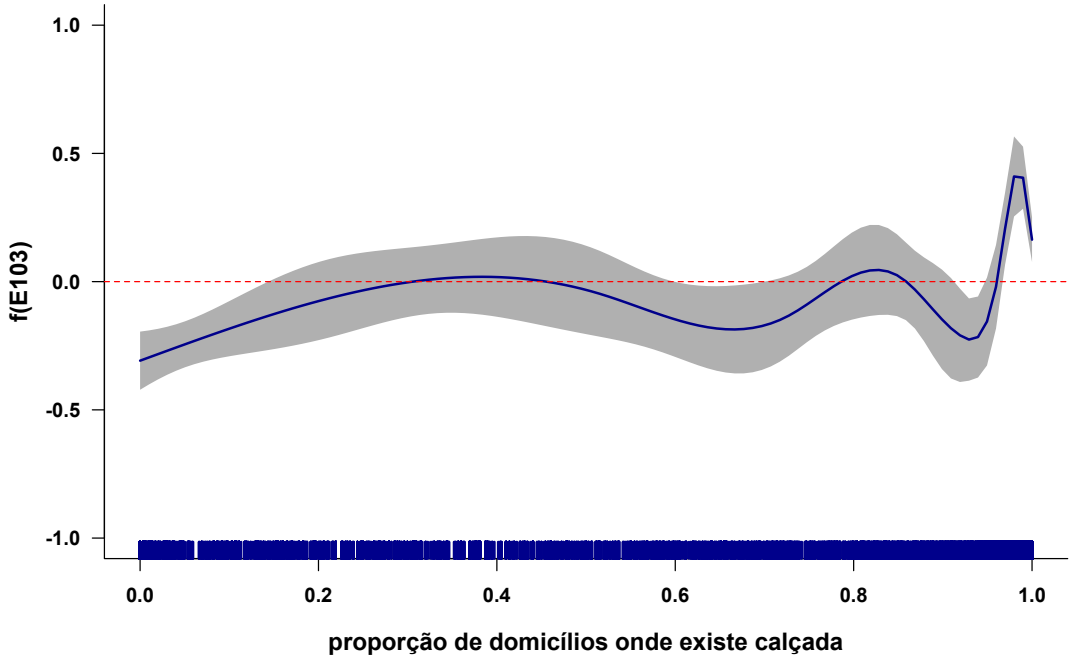


Figura 24: Gráfico do efeito parcial da variável $E103$.

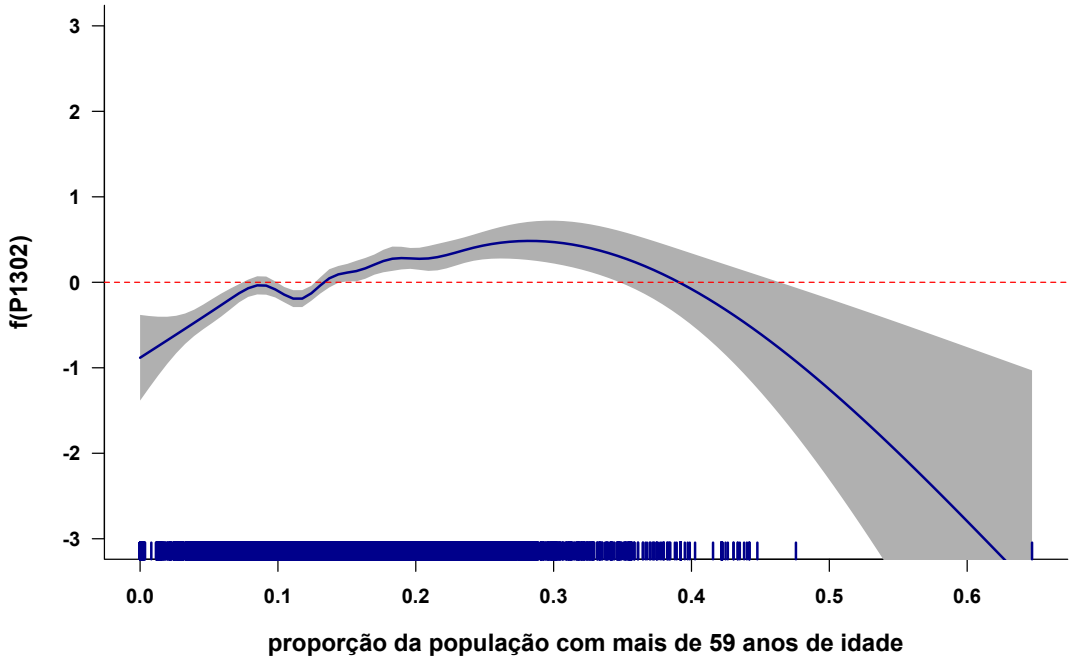


Figura 25: Gráfico do efeito parcial da variável $P1302$.

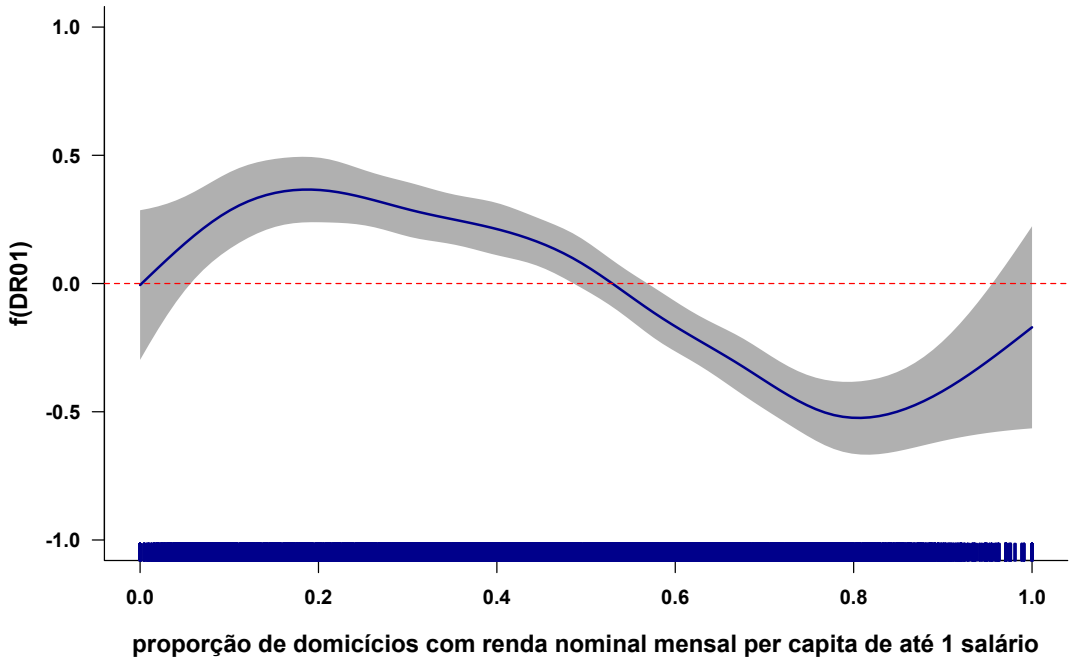


Figura 26: Gráfico do efeito parcial da variável $DR01$.

modelo, percebe-se que proporções menores do que 53% possuem efeito negativo e proporções maiores do que esse valor, positivo. Ou seja, regiões onde há menor número de pessoas mais pobres são mais suscetíveis à ocorrência de roubos e regiões onde há maior volume de população pobre, menos. As proporções entre 0 e $0,6$ e acima de $0,95$ não são significativas.

Em regiões mais povoadas, é provável que haja um número maior de roubos e, assim, que a probabilidade de um roubo acontecer em tal região seja maior. A Figura 27, que mostra o efeito parcial da população média por célula, sustenta essa afirmação ao indicar uma curva que cresce à medida que a média cresce, com pico em 650 pessoas, e permanece estabilizada a partir desse valor, até se tornar não significativa a partir de 1810. A variável $PQ1$ possui efeito negativo do valor mínimo da média até 140 e após esse valor, o efeito torna-se positivo.

Para que fosse efetuada uma interpolação usando o ajuste do modelo proposto, foram gerados 36.824 pontos em uma grade regular dentro da região das 20.667 quadrículas no mapa do município do Rio de Janeiro. O resultado é exibido na Figura 28. Tons avermelhados indicam uma maior probabilidade de ocorrência de roubo e tons mais azulados, menores probabilidades. Percebe-se maiores concentrações de tons avermelhados pela zona norte. As maiores probabilidades de roubo da zona oeste se dão na região de Magalhães Bastos, Realengo e Padre Miguel, locais próximos à zona norte. Também

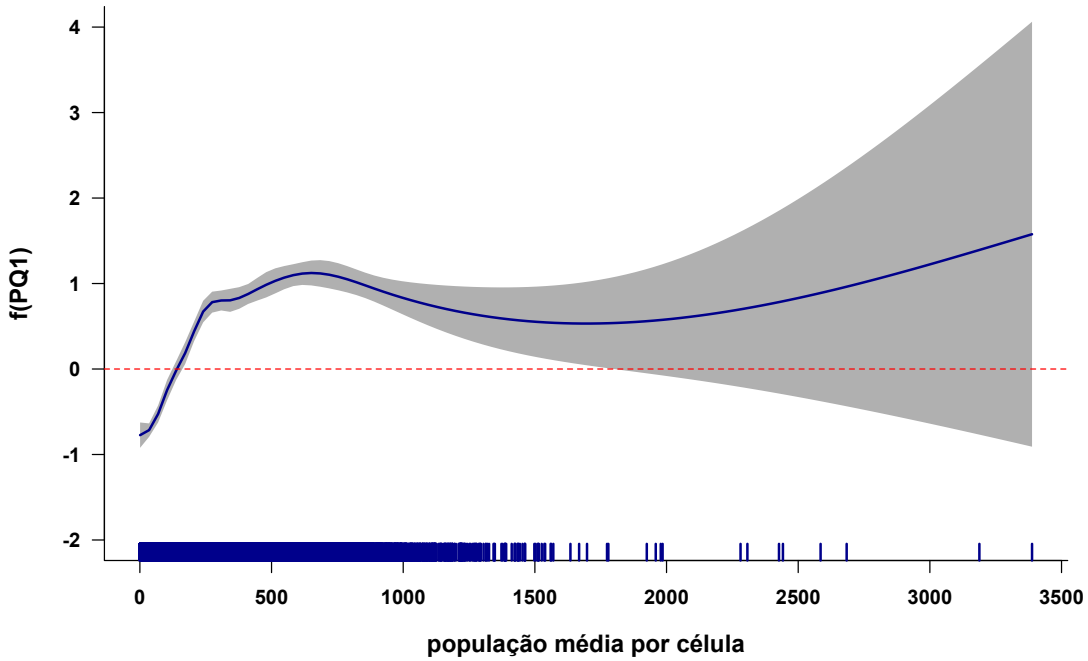


Figura 27: Gráfico do efeito parcial da variável $DR01$.

há concentrações de tons avermelhados pela zona sul e região do centro da cidade. Das probabilidades interpoladas, metade delas se encontram abaixo de $34,08\%$, porém 25% dos valores possuem probabilidade bastante alta, entre $78,85\%$ e $99,81\%$, o que explica a quantidade de tons avermelhados vistos na Figura 28.

O maior valor de probabilidade interpolada se dá bem próximo à Praça da Bandeira, entre a Avenida Oswaldo Aranha e Rua Paulo Fernandes, com $99,81\%$. O menor valor interpolado se dá em Guaratiba, próximo à Estrada Paiva Muniz, área do Parque Estadual da Pedra Branca, com $0,13\%$. A Tabela 5 mostra o quanto cada função suavizada contribui para que se chegue nesses valores de probabilidade. A covariável que tem maior influência em ambos os resultados é a distância até a via mais próxima.

Comparando-se três pontos no município, o primeiro no Leblon, na esquina das ruas Dias Ferreira e Rainha Guilhermina, o segundo no Centro, no Largo do Paço, próximo à Praça XV e o terceiro em Campo Grande, em frente ao Hospital Municipal Rocha Faria, tem-se que a chance de ocorrer roubo no ponto do Leblon em relação ao de Campo Grande é $76,84\%$ menor. Comparando-se o ponto do Leblon com o do Centro, a chance de ocorrer roubo no Leblon é $73,71\%$ menor. Comparando-se os pontos de Campo Grande e Centro, as chances de ocorrer roubo em Campo Grande é 14% maior.

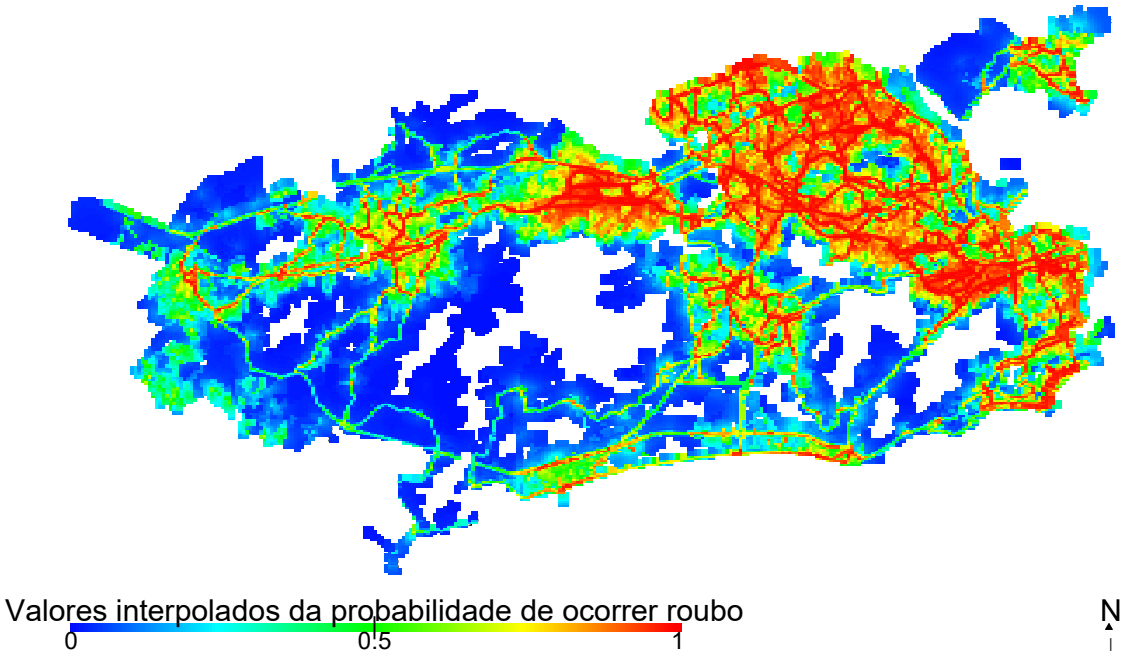


Figura 28: Mapa dos valores interpolados para os 36.824 pontos.

Tabela 5: Contribuição dos efeitos parciais de cada função suavizada das covariáveis para o valor interpolado

função suavizada	Valor Máximo	Valor Mínimo
$f_1(dist\_via)$	2.4648	-1.8424
$f_2(dist\_fav)$	0.1520	-0.5042
$f_3(dist\_esc)$	0.1524	-0.7626
$f_4(dist\_trem)$	0.6962	-0.5828
$f_5(D106)$	1.1200	-0.0997
$f_6(E103)$	0.1636	-0.3083
$f_7(P1302)$	0.4413	-0.1265
$f_8(DR01)$	0.3601	-0.3700
$f_9(PQ1)$	1.0853	-0.7666
$f_{10}(lon, lat)$	0.1336	-0.7960

4 Conclusão

Os roubos são crimes que causam impacto grande na vida da sociedade carioca e brasileira em geral, refletindo tanto em perdas materiais quanto de bem-estar para os cidadãos. Foi mostrado, aqui, que em $42,78\%$ das quadrículas do município do Rio ocorreu roubo no ano de 2019 (Figura 14). Daí vem a importância de agregar informações sobre esse tipo de atividade criminosa, para tentar oferecer suporte à tomada de decisões, ou causar reflexão sobre por onde e como pode ser melhor combatida. Assim, a tentativa de identificar, compreender e descrever o padrão de comportamento de dados de roubos na cidade do Rio de Janeiro foi a principal motivação deste trabalho. A incorporação de informações espaciais ao modelo aditivo generalizado proposto pelo estudo foi feita e se mostrou importante, dado que a longitude e latitude das quadrículas se mostrou significativa. Os dados sociodemográficos do Censo de 2010 e as geolocalizações de lugares de interesse também agregaram valor às análises.

Para realização deste estudo, foi necessário pesquisar e determinar de forma detalhada as quadrículas habitadas, para que as não habitadas fizessem parte do banco de dados na quantidade mínima possível para que não houvesse confusão nos efeitos do modelo.

Com a análise descritiva, além de determinar a quantidade de quadrículas em que houve roubo e quantos roubos houve em média por quadrícula, foi possível perceber que houve menor incidência de roubo nas quadrículas pertencentes à zona oeste da cidade (Figura 14), mas isso também está relacionado à menor população média por célula (Figura 18(d)). Algumas variáveis não aparentaram trazer quaisquer informações importantes por terem a grande maioria de seus valores em um extremo ou não demonstrarem padrão espacial algum e esse fato foi confirmado nos testes para avaliação das variáveis a entrar no modelo.

Os resultados do modelo mostraram que o fato de estar muito próximo à uma via rápida faz com que a probabilidade de ocorrência de roubo suba bastante (Figura 19), concordando com os resultados de estudos comentados na Seção 1.2. Locais com maiores

proporções de idosos e menores proporções de moradores do sexo masculino também fazem a probabilidade de roubo aumentar, assim como a proximidade de locais com grande fluxo de pessoas como escolas e estações de trem. Essas informações oferecem suporte para indicar a direção e o foco que as autoridades devem tomar para melhorar e otimizar o combate desse crime na cidade. De uma forma geral o mapa das probabilidades de ocorrência de roubos interpoladas para o município (Figura 28) apresenta uma quantidade maior de probabilidades altas, fato que é preocupante.

Para estudos futuros, seria importante que houvesse acesso à informações melhores organizadas e de maneira mais fácil sobre outros possíveis tipos de variáveis como locais públicos para lazer ou comércio. Agregações em menores níveis de área para variáveis socioeconômicas e educacionais, além das contidas no censo, também trariam importantes informações para estudos com alto nível de detalhamento e resolução espacial. Outro ponto possível de melhora dos resultados obtidos neste trabalho é o agrupamento por espaços de tempo menores.

Foi decidido que para este trabalho o tempo não seria um fator de análise, para que houvesse um maior equilíbrio entre complexidade e tempo de preparação do mesmo. Estudos futuros devem incorporar, de alguma forma, o tempo como covariável. Uma análise espaço-temporal seria um próximo passo, que provavelmente melhoraria os resultados e abriria a possibilidade de se investigar outros fatores, como se em um período do ano há mais roubos do que em outros, ou de se associar mais indicadores, como por exemplo, momentos do dia em que ocorrem os roubos ou, até mesmo, se há algum evento ou aglomeração naquele local e tempo.

Referências

BEATO, C. C. Fontes de dados policiais em estudos criminológicos: limites e potenciais. In: FÓRUM DE DEBATES: CRIMINALIDADE, VIOLÊNCIA E SEGURANÇA PÚBLICA NO BRASIL: UMA DISCUSSÃO SOBRE AS BASES DE DADOS E QUESTÕES METODOLÓGICAS, 1º Encontro., 2000, Brasília. *Anais*. [S.l.]: Instituto de Pesquisa Econômica Aplicada (IPEA), 2000.

CERQUEIRA, D. R. d. C. et al. Atlas da violência 2019. Instituto de Pesquisa Econômica Aplicada (IPEA), 2019.

DOBSON, A. J.; BARNETT, A. G. *An introduction to generalized linear models*. [S.l.]: Chapman and Hall/CRC, 2018.

GARETH, J. et al. *An introduction to statistical learning: with applications in R*. [S.l.]: Spinger, 2013.

GONÇALVES, L. d. L. Estruturação, representação e análise de dados criminais: o caso da parceria entre o isp e a drfc. *Cadernos de Segurança Pública*, Instituto de Segurança Pública/RJ (ISP), n. 13, Dezembro 2021.

HASTIE, T.; TIBSHIRANI, R. *Generalized additive models (Monographs on statistics and applied probability 43)*. [S.l.]: Chapman & Hall, 1990.

LIZZI, E. A. da S. et al. Homicides of black people in brazil: A study of different regions, using generalized additive regression models-with a geo-spatial component. *Geospatial health*, v. 16, n. 1, 2021.

MAHFOUD, M.; BHULAI, S.; MEI, R. van der. Crime forecasting in small city blocks using a general additive spatio-temporal model. *International Journal on Advances in Security*, v. 11, n. 3 & 4, p. 214–222, 2018.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to linear regression analysis*. 5. ed. [S.l.]: John Wiley & Sons, 2012.

OpenStreetMap Foundation. *Licence/Attribution Guidelines — OpenStreetMap Foundation*. 2022. Acesso em: 1 dezembro de 2022. Disponível em: ⟨https://wiki.osmfoundation.org/w/index.php?title=Licence/Attribution_Guidelines&oldid=9351⟩.

OSORIO, M.; VERSIANI, M. H.; VEIGA, L. A. d. O círculo vicioso de violência no Rio de Janeiro. *Jornal dos Economistas*, Conselho Regional de Economia/RJ (CORECON), n. 345, Maio 2018.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2022. Disponível em: ⟨http://www.R-project.org/⟩.

WANG, X.; BROWN, D. E. The spatio-temporal modeling for criminal incidents. *Security Informatics*, SpringerOpen, v. 1, n. 1, p. 1–17, 2012.

WOOD, S. N. *Generalized additive models: an introduction with R*. 2. ed. [S.l.]: chapman and hall/CRC, 2017.